UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA (UESB)
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA (PPGLIN)

JOICE MALTA SANTOS

UMA PROPOSTA DE WORKFLOW PARA CONSTRUÇÃO DE CORPUS DIGITAL EM LÍNGUA DE SINAIS

VITÓRIA DA CONQUISTA - BA
2024

JOICE MALTA SANTOS

UMA PROPOSTA DE WORKFLOW PARA CONSTRUÇÃO DE CORPUS DIGITAL EM LÍNGUA DE SINAIS

Dissertação apresentada ao Programa de Pós-Graduação em Linguística (PPGLin), da Universidade Estadual do Sudoeste da Bahia (UESB), como requisito parcial e obrigatório para obtenção do título de Mestre em Linguística.

Área de Concentração: Linguística

Linha de Pesquisa: Descrição e Análise de Línguas Naturais

Orientadora: Prof.ª Dr.ª Cristiane Namiuti

Coorientadora: Prof.ª Dr.ª Adriana Stella Cardoso Lessa-de-Oliveira

VITÓRIA DA CONQUISTA - BA

2024

Santos, Joice Malta.
S236p Uma proposta de *workflow* para construção de *corpus* digital em língua de sinais. / Joice Malta Santos; orientadora: Cristiane dos Santos Namiuti. – Vitória da Conquista, 2024.
141f.

Dissertação (mestrado – Programa de Pós-Graduação em Linguística) – Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2024.
Inclui referência F. 107 – 111.

1. Língua de Sinais. 2. Construção de corpora. 3. Método Lapelinc. 4. Escrita Sel. I. Namiuti, Cristiane dos Santos (orientadora). II. Universidade Estadual do Sudoeste da Bahia, Programa de Pós-Graduação em Linguística. T. III

CDD: 419

Catalogação na fonte: *Karolyne Alcântara Profeta — CRB 5/2134*
UESB – *Campus* Vitória da Conquista – BA

Título em inglês: A workflow proposal for construction digital corpus in sign language.

Palavras-chave em inglês: Sign language. Corpora construction. Lapelinc method. Sel writing.

Área de concentração: Linguística.

Titulação: Mestre em Linguística.

Banca examinadora: Prof.ª Dr.ª Cristiane Namiuti (Presidente-Orientadora, UESB); Prof.ª Dr.ª Adriana Stella Cardoso Lessa-de-Oliveira (Coorientadora, UESB); Prof. Dr. Jorge Viana Santos (UESB); Prof. Dr. João Paulo Lazzarini (UFBA).

Data da defesa: 19 de novembro de 2024.

Programa de Pós-Graduação: Programa de Pós-Graduação em Linguística.

Orcid ID: https://orcid.org/0000-0003-0569-7153

Lattes ID: https://lattes.cnpq.br/0328657373568342

JOICE MALTA SANTOS

UMA PROPOSTA DE WORKFLOW PARA CONSTRUÇÃO DE CORPUS DIGITAL EM LÍNGUA DE SINAIS

Dissertação apresentada ao Programa de Pós-Graduação em Linguística, da Universidade Estadual do Sudoeste da Bahia, como requisito parcial e obrigatório para a obtenção do título de Mestre em Linguística.

Data da aprovação: 19 de novembro de 2024.

Banca Examinadora:

Profa. Dra. Cristiane dos Santos Namiuti
Instituição: UESB – Presidente-Orientadora
Ass.: Cristiane Namiuti

Profa. Dra. Adriana Stella Cardoso Lessa de Oliveira
Instituição: UESB – Coorientadora
Ass.: Documento assinado digitalmente gov.br ADRIANA STELLA CARDOSO LESSA DE OLIVEIRA Data: 03/12/2024 10:15:44-0300 Verifique em https://validar.iti.gov.br

Prof. Dr. Jorge Viana Santos
Instituição: UESB – Membro Titular
Ass.: Documento assinado digitalmente gov.br JORGE VIANA SANTOS Data: 23/11/2024 18:41:46-0300 Verifique em https://validar.iti.gov.br

Prof. Dr. João Paulo Lazzarini Cyrino
Instituição: UFBA – Membro Titular
Ass.: Documento assinado digitalmente gov.br JOAO PAULO LAZZARINI CYRINO Data: 19/11/2024 14:23:53-0300 Verifique em https://validar.iti.gov.br

AGRADECIMENTOS

O ato de agradecer é intenso e plausível. Ele nos permite reconhecer nos detalhes os prazeres da vida e valorizar as pessoas que contribuem para que nossa caminhada seja mais fluida e feliz. É com essa certeza que deixo aqui meus agradecimentos a todos aqueles que de forma direta ou indireta colaboraram para que a realização dessa pesquisa fosse possível.

Agradeço primeiramente a Deus, aquele que me deu a vida e que tem me sustentado durante todos os meus dias.

À Universidade Estadual do Sudoeste da Bahia (UESB) e ao Programa de Pós-Graduação em Linguística (PPGLin), pela oportunidade de realização da minha formação em nível de mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[1], pelo apoio e financiamento das atividades do PPGLin da UESB.

À Prof.ª Dr.ª Cristiane Namiuti, pelos muitos conhecimentos adquiridos a partir das suas orientações, pela competência, paciência e compreensão. À Prof.ª Dr.ª Adriana Lessa-de-Oliveira, pela coorientação e contribuição neste trabalho. Vocês são excelentes profissionais, de fato, exemplos a serem seguidos.

Ao Prof. Dr. Jorge Viana Santos e ao Prof. Dr. João Paulo Lazzarini, por aceitarem compor as bancas de qualificação e defesa e pelas valiosas contribuições dadas à minha pesquisa.

À minha família, em especial aos meus pais Jesus e Edna, à minha irmã Nátally e ao meu namorado Bruno, por sempre ser o meu apoio, encorajamento e por entenderem os momentos em que precisei me fazer ausente em decorrência dos estudos. Obrigada por acreditarem em mim.

Aos meus amigos, com os quais compartilhei minhas pequenas conquistas e eles se alegraram comigo.

[1] Conforme Portaria CAPES nº 206, de 4 de setembro de 2018, “O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001”.

RESUMO

Os corpora de línguas de sinais disponíveis atualmente em pesquisas linguísticas e em sites para acesso livre são constituídos por um módulo de gravação feita em vídeo, pois os dados pertencem a uma língua de modalidade gesto-visual e outro módulo de "transcrição" dos dados por meio de glosas que utilizam a escrita de uma língua oral, como o Português ou o Inglês. No entanto, esse método comumente utilizado para descrever e estudar as propriedades gramaticais das línguas de sinais apresenta um problema: a falta de um módulo que desempenhe, de fato, a função da transcrição, que é a de realizar a associação entre forma e sentido por meio de um sistema de anotação escrita do sinal. Isso porque, ao analisar essas "transcrições" por glosa, percebe-se que elas exercem, na verdade, um papel de tradução e não de transcrição, uma vez que não se utiliza um sistema de escrita equivalente à escrita fonética para transcrever os sinais; as traduções não são representações do sinal, pois não consideram sua articulação na anotação, são apenas traduções para a língua oral dos sentidos do que foi sinalizado no vídeo. Esse fato se apresenta como um problema para os estudos linguísticos, pois pode afetar a análise e descrição das propriedades linguísticas, interferindo na formulação das hipóteses sobre a estrutura gramatical da língua de sinais por parte do pesquisador. Uma proposta que levantamos como enfrentamento a esse problema é a utilização de um sistema de escrita para língua de sinais que represente a articulação fonética do sinal para compor o módulo de transcrição nas iniciativas de construção de corpora para essas línguas. O sistema que melhor atendeu ao objetivo de representar linearmente a articulação fonética de sinais mano-visuais de línguas sinalizadas foi o Sistema de Escrita para Língua de Sinais (SEL) (Lessa-de-Oliveira, 2023). Nesse contexto, essa dissertação teve como objetivo principal: (a) realizar um levantamento de como os estudos da área da Linguística trabalham com os dados das línguas de sinais (especificamente da Libras) e como são constituídos os corpora que fundamentam esses estudos; (b) avaliar os limites e as possibilidades das iniciativas de construção de corpora para línguas de sinais encontradas; e (c) propor um fluxo de trabalho, um workflow, para construção de corpora de línguas de sinais, que atenda todas as etapas iniciais de anotação incluindo a possibilidade de transcrições fonéticas e/ou fonológicas, que siga as diretrizes de construção de corpora para línguas oro-auditivas no que se refere às possibilidades de anotação e que faça reuso das tecnologias já existentes. Para isso, foi executada uma metodologia de pesquisa em nosso trabalho que se caracteriza como pesquisa aplicada, na qual seguimos as etapas: (i) realização de um estudo para conhecimento do estado da arte e levantamento de corpora de língua de

sinais observando seus alcances, possibilidades e limites; (ii) levantamento de requisitos importantes para anotação e controle na produção de corpora de língua de sinais; (iii) definição dos metadados que podem guiar a catalogação dos dados, os quais devem ser guardados em uma estrutura de arquivo como corpus cru (Santos; Namiuti, 2019); (iv) construção de um modelo teórico a partir do workflow proposto para a construção de corpora de língua de sinais que contempla o módulo de transcrição e tradução com exemplificação da anotação a partir de dados de Libras selecionados nas fontes da pesquisa; (v) realização de análise comparativa entre o modelo construído a partir do workflow proposto e os corpora construídos com a utilização de outros esquemas para investigações da língua Libras. A partir disso, foi possível demonstrar que: (1) uma iniciativa de construção de corpora para línguas de sinais que siga padrões de anotação sistematizados e padronizados é necessária para que pesquisas na área da Linguística de línguas sinalizadas (mano-visuais), no caso do Brasil a Libras, sejam bem fundamentadas; e (2) a escrita Sel se caracteriza como uma proposta plausível para compor o módulo de transcrição nessa iniciativa.

PALAVRAS-CHAVE

Língua de Sinais; Construção de corpora; Método Lapelinc; Escrita Sel.

ABSTRACT

The sign language corpora currently available in linguistic research and on open-access websites consist of a video recording module, since the data belongs to a language with a gesture-visual modality, and another module for "transcribing" the data through glosses that use the writing of an oral language, such as Portuguese or English. However, this method commonly used to describe and study the grammatical properties of sign languages presents a problem: the lack of a module that actually performs the function of transcription, which is to associate form and meaning through a system of written annotation of the sign. This is because, when analyzing these "transcriptions" by gloss, it is clear that they actually play the role of translation and not transcription, since a writing system equivalent to phonetic writing is not used to transcribe the signs; the translations are not representations of the sign, since they do not consider its articulation in the annotation; they are merely translations into the oral language of the meanings of what was signaled in the video. This fact presents itself as a problem for linguistic studies because it can affect the analysis and description of linguistic properties, interfering with the formulation of hypotheses about the grammatical structure of sign language by the researcher. One proposal that we put forward to address this problem is the use of a writing system for sign language that represents the phonetic articulation of the sign to compose the transcription module in initiatives to build corpora for these languages. The system that best met the objective of linearly representing the phonetic articulation of mano-visual signs of signed languages was the Writing System for Sign Language (SEL) (Lessa-de-Oliveira, 2023). In this context, this dissertation had as its main objective: (a) to carry out a survey of how studies in the field of Linguistics work with data from sign languages (specifically Libras) and how the corpora that support these studies are constituted; (b) to evaluate the limits and possibilities of the initiatives to build corpora for sign languages found; and (c) propose a workflow for building sign language corpora that meets all the initial stages of annotation, including the possibility of phonetic and/or phonological transcriptions, that follows the guidelines for building corpora for oral-auditory languages with regard to annotation possibilities, and that reuses existing technologies. To this end, a research methodology was implemented in our work that is characterized as applied research, in which we followed the steps: (i) conducting a study to understand the state of the art and surveying sign language corpora, observing their scope, possibilities, and limits; (ii) surveying important requirements for annotation and control in the production of sign language corpora; (iii) defining the metadata that can guide the cataloging of data, which must be saved in a file

structure as a raw corpus (Santos; Namiuti, 2019); (iv) construction of a theoretical model based on the proposed workflow for constructing sign language corpora that includes the transcription and translation module with annotation examples based on Libras data selected from the research sources; (v) performance of a comparative analysis between the model constructed based on the proposed workflow and the corpora constructed using other schemes for Libras language investigations. From this, it was possible to demonstrate that: (1) an initiative to construct corpora for sign languages that follows systematized and standardized annotation patterns is necessary for research in the area of Linguistics of signed languages (manual-visual), in the case of Brazil Libras, to be well-founded; and (2) the Sel script is characterized as a plausible proposal to compose the transcription module in this initiative.

KEYWORDS

Sign language; Corpora construction; Lapelinc method; Sel writing.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AD	Análise Descritiva
AME	Aparato de Metadados Estruturados
AT	Análise Topográfica
CD	Configuração de Dedos
CM	Configuração de Mão
CMov	Composição entre os Movimentos das duas mãos
CompM	Composição de Mãos
CORDIAL SIN	Corpus Kadiwéu, do Corpus Dialetal para o Estudo da Sintaxe
Corpus DOVIC	Corpus de Documentos Oitocentistas de Vitória da Conquista
CRPC	Corpus de Referência do Português Contemporâneo
CTB	Corpus Tycho Brahe
CV	Catálogo Visual
D/CD	Dedo/Combinação de Dedo
DDI	Documento Digital Imagem
DDT	Documento Digital Texto
DF	Documento Físico
DMov	Direção do Movimento
DOP	Dossiê de Observações Pertinentes
ELAN	Eudico Language Annotator
ELiS	A Escrita de Língua de Sinais
EM	Eixo de Posição da Mão
ENM	Expressões Não Manuais
ExpF	Expressão Facial
FCC	Fotografia Cientificamente Controlada
GI	Grupo de Aquisição na Infância
GO	Grupo Ouvintes
GPI	Grupo de Aquisição Pós-Infância
L	Locação
LAPELINC	Laboratório de Pesquisa em Linguística de Corpus
Libras	Língua Brasileira de Sinais
LLIC	Laboratório de Linguagem, Interação e Cognição
L1	Primeira Língua

L2	Segunda Língua
M	Mão
MLMov	Mão, Locação e Movimento
Mov	Movimentos
NILC	Núcleo Interinstitucional de Linguística Computacional
OP	Direção ou Orientação da Palma
OrT/PPC	Ordenamento de Toque/Proximidade em Partes do Corpo
PA	Ponto de Articulação
PC	Parte do Corpo |PC|
PDM	Posicionamento das Duas Mãos
PE	Português Europeu
PLN	Processamento de Linguagem Natural
PMov	Plano de Movimento
Sel	Sistema de Escrita da Libras
SEU	Survey of English Usage
SW	SignWriting
TMovD	Tipo de Movimento de Dedo
TMovM	Tipo de Movimento de Mão
T/PM	Toque/Proximidade na Mão
T/PPC	Toque/Proximidade em Parte do Corpo
VisoGrafia	Escrita Visogramada das Língua de Sinais
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

Em consonância com MacCarthy e O'Keeffe (2010), os corpora possuem extrema relevância em um trabalho de investigação linguística, uma vez que, com a quantidade extensa de dados de língua depositados neles, torna-se possível a verificação de ocorrências, viabilizando a explicação, pelo pesquisador, dos acontecimentos linguísticos. Nesse caso, a tarefa do pesquisador se aplica à construção de metodologias confiáveis para analisar, descrever e explicar tais acontecimentos. Sendo assim, é a partir da precisa disposição dos dados presentes nos corpora que as pesquisas se fazem capazes de serem realizadas com êxito.

Desse modo, dada a sua acentuada importância, é correto afirmar que é imprescindível que as iniciativas de construções de corpora linguísticos sejam bem elaboradas a ponto de exercerem com propriedade sua função. Isso ocorre muito satisfatoriamente com as línguas orais (oro-auditivas), entretanto, no que diz respeito às línguas de sinais (mano-visuais), o mesmo não acontece, pois, após estudos e levantamentos, foi possível verificar que as iniciativas existentes de construção de corpora para línguas de sinais são insuficientes para a elaboração de estratégias para construção de corpora que de fato atendam às exigências da área da Linguística e que se façam coerentes para uma análise linguística precisa.

Quadros (2016) destaca que o estado da iniciativa de construção de corpora para línguas de sinais se configura de maneira distinta das iniciativas de línguas orais, uma vez que aquelas são línguas que se realizam pela modalidade gesto-visual e são, ainda, ágrafas[2] – o fato de se configurarem de maneira distinta não anula a necessidade de serem bem sistematizadas a ponto de cumprirem satisfatoriamente sua função. Sendo assim, as produções, nas iniciativas de construção de corpora para línguas de sinais, são realizadas por meio de vídeo devido, justamente, ao fato de se tratarem de dados de uma língua de modalidade gesto-visual e as "transcrições" são na verdade traduções glosadas a partir da escrita de uma língua de modalidade oral estabelecida para isso (o Português ou o Inglês, por exemplo), pois as línguas de sinais ainda não possuem um sistema de escrita consolidado, amplamente utilizado e que reflita a articulação/produção do sinal gesto/visual[3]. No entanto,

[2] Apesar de haver propostas de sistemas de escrita para as línguas de sinais, elas ainda são consideradas ágrafas porque nenhum desses sistemas está, até o presente momento, de fato em uso efetivo pela comunidade surda e pelos falantes de línguas de sinais no dia a dia; são sistemas utilizados, por hora, apenas na academia para fins de pesquisa.

[3] Vale a pena refletir que essas duas características – se realizar pela modalidade gesto-visual e ser ágrafa – podem estar intrinsecamente ligadas, pois a dificuldade de grafar as línguas de sinais talvez se configure pela cultura de escrita que temos para as línguas da modalidade oral em que é recorrente uma escrita alfabética muito ligada à articulação oral, isto é, uma escrita que representa uma

este método amplamente utilizado como solução para a descrição e estudo das propriedades gramaticais das línguas de sinais traz consigo um problema intrínseco que pode interferir na interpretação e análise das propriedades linguísticas das línguas de sinais, uma vez que a articulação do sinal da língua alvo de descrição e análise não é anotada por meio de transcrição 'fonética'/'fonológica'[4], ou seja, não é adotado um sistema de escrita equivalente à escrita fonética para transcrever o sinal.

Para assimilar o problema do módulo de transcrição, é preciso questionar e ao mesmo tempo esclarecer sobre um ponto que defendemos neste trabalho: a transcrição de uma determinada língua tem o objetivo de anotar a articulação das formas e estruturas da língua transcrita (os fonemas e morfemas). Na língua portuguesa, isso é alcançado pela transcrição fonética nos corpora que têm a oralidade como objeto de pesquisa, algo que, claramente não acontece em corpora de línguas de sinais, pois, pela falta de um sistema de escrita fonético/articulatório, há a supressão dessa etapa de descrição/anotação na construção de corpora para estudo linguístico ou substituição dele por um módulo de transcrição feito exclusivamente pelo pareamento de imagens glosadas em uma língua pré-estabelecida para a descrição. Assim, os estudos que tem como alvo a descrição linguístico-gramatical de línguas de sinais, como, por exemplo, da Libras, partem de glosas, ou seja, de traduções, para a elaboração de hipóteses sobre a língua, o que pode, em algum grau, interferir nas conclusões que se faz sobre a língua alvo do estudo.[5]

O fato da glosa não ser e nem equivaler de fato a uma transcrição das línguas de sinais, mas sim a uma tradução, acarreta no problema da ausência de transcrição nos corpora de

---

articulação oral produzida pelo aparato fonador, no espaço interno do corpo, e percebida pela audição, característica das línguas oro-auditivas. A articulação das línguas de sinais é produzida pelas mãos num espaço externo do corpo e percebida pela visão, outro sentido, por ter essas características, tais línguas são denominadas mano-visuais ou gesto-visuais. Sendo assim, em termos abstratos, essas línguas comungam de um mesmo sistema que é a existência de uma articulação e uma percepção, mas, em termos concretos, eles se dão de formas distintas: (i) o primeiro por meio de um sistema articulatório com um aparato fonador que produz sinais acústicos (orais) e um sistema de percepção do sinal que é auditivo; (ii) o segundo, por meio de um sistema articulatório com um aparato gestual que se utiliza das mãos e partes do corpo para a articulação do sinal manual e um sistema de percepção do sinal que é visual.

[4] Nas línguas orais se utiliza essa transcrição para se anotar a articulação do sinal acústico (oral). Já nas línguas de sinais essa transcrição é necessária para se anotar a articulação do sinal gestual (manual). Essa associação é possível uma vez que as línguas de sinais também se classificam como línguas naturais e por essa razão também possuem estrutura fonética e fonológica.

[5] A utilização de escrita alfabética como base de transcrição de dados/textos de línguas orais não é equivalente às traduções glosadas que são utilizadas nos estudos de línguas sinalizadas, pois trata-se de um sistema de escrita baseado na fonética-fonologia das línguas orais. A escrita alfabética, apesar de ser insuficiente para retratar certos fenômenos da língua, é suficiente para descrever uma forma, um sinal de base oral e estabelecer a relação entre este sinal acústico e um sentido.

línguas de sinais, o que traz necessariamente uma limitação ao corpus, uma vez que ter a possibilidade de transcrever é um requisito fundamental para construir um corpus anotado de língua natural, porque o que se almeja no final do processo de construção do corpus é ter um registro em texto da língua alvo da descrição/anotação. A falta da possibilidade de transcrição também tem uma consequência direta na falta de ferramentas para construção de hipóteses sobre a língua. Isso tudo não quer dizer que a tradução por glosa não é importante, ela é importante, pois dá acesso à porção de sentido do significante; mas sim quer dizer que ela não é suficiente para formular hipóteses sobre a língua, pois é necessário a transcrição para se ter acesso ao significante, acesso este que possa ser lido e anotado também pelo computador. Por isso, é necessário se ter, nas iniciativas de construção de corpora para línguas de sinais, um sistema de transcrição para esta língua que cumpra, estritamente, com seu papel de representar graficamente o sinal.

Posto isso, uma proposta que aventamos para responder ao problema da não existência de um módulo de transcrição do sinal nas atuais iniciativas de construção de construção de corpora para línguas de sinais é a utilização do Sistema de Escrita para Língua de Sinais, a escrita SEL, desenvolvido pela Prof.ª Dr.ª Adriana Stella Cardoso Lessa-de-Oliveira (2012, 2019 e 2023). A escrita Sel tem o objetivo e ao mesmo tempo o desafio de representar linearmente a articulação do sinal de línguas de sinais com foco na Língua Brasileira de Sinais (Libras) a fim de incluir os falantes/sinalizantes dessa língua no mundo letrado. Para isso, Lessa-de-Oliveira (2023) explica que essa proposta é possível porque os sinais são formados por unidades constituídas por três elementos: mão, locação e movimento, denominando, assim, a unidade MLMov. Isto é, na escrita SEL, a representação das unidades MLMov marca cada traço da configuração tridimensional do sinal. Nesse sentido, o sinal ilustrado na figura 1 abaixo, que pode ser traduzido para o Português como o nome “trabalho” ou como o verbo “trabalhar”, pode ser escrito pela SEL como está ilustrado na figura 2 mais abaixo:

Figura 1 – Sinal em Libras equivalente ao sentido de “Trabalh(o)/(ar)” em Português

Fonte: https://bit.ly/3mpGjDR

Figura 2 – Escrita do Sinal em Libras equivalente ao sentido de “Trabalh(o)/(ar)” em Português, sistema de escrita SEL

Fonte: Elaboração própria com base nas regras da Sel apresentada em Lessa-de-Oliveira (2023)

Lessa-de-Oliveira (2023) postula que a escrita Sel representa as unidades de elementos que formam o sinal por meio de letras e diacríticos, ordenando da esquerda para a direita, com a sequência de mão, locação e movimento. Percebemos, assim, que, com a escrita SEL, é possível se alcançar o objetivo de uma transcrição, o qual já foi mencionado anteriormente, que é o de anotar a articulação do sinal, uma vez que olhando para a escrita Sel acima exposta do sinal “trabalho/trabalhar”, tendo como pré-requisito o conhecimento das regras de funcionamento dessa escrita, temos as diretrizes necessárias para conseguir reproduzir, na modalidade gesto-visual, este sinal. Assim, com a Sel é possível se fazer a associação entre forma e sentido em registro escrito dos sinais, isto é, ter um texto anotado que represente precisamente a articulação do sinal, tornando viável um módulo de transcrição que contemple o sistema articulatório perceptual da língua de sinais alvo. Entendemos, portanto, que a transcrição por escrita Sel, que é padronizada, num corpus de língua de sinais é equivalente à transcrição fonética/fonológica num corpus de língua oral.

Atendido o requisito da etapa de transcrição para a construção de um corpus anotado em meio digital, através da utilização do sistema Sel, temos a possibilidade de pensar em um sistema para a elaboração de corpus anotado para línguas de sinais. Nesse sentido, nossa pesquisa centrou-se no levantamento das metodologias de construção de corpora para língua de sinais e para línguas orais buscando identificar limites e soluções nas iniciativas existentes para corpora anotados para estudos linguísticos e, a partir disso, pensar uma proposta de reuso de tecnologias de construção de corpora em uma metodologia que inclua o módulo de transcrição da língua de sinais que passa por uma codificação escrita da articulação do sinal nos moldes de uma transcrição fonética para as línguas orais.

Defendemos que a metodologia desenvolvida por Santos e Namiuti (2019), o método Lapelinc de construção de corpora, que em sua essência e filosofia prevê controle científico para garantir o resgate automático de informações fidedignas à fonte original do texto e que atendam todas as diferentes possibilidades de interesses acadêmico-científico-social, pode ser utilizada com algum grau de adaptação, para a construção de corpora de línguas de sinais. Para a elaboração de nossa proposta consideramos as etapas e subprodutos propostos pelo

LAPELINC Framework (Costa, Santos e Namiuti, 2021). Essa nossa proposta se justifica porque o LAPELINC Framework e está sendo desenvolvido para instruir iniciativas de construção de corpora digitais para línguas naturais, para se configurar como um padrão de referências para essas línguas e, sendo assim, a línguas de sinais como língua natural se enquadra indubitavelmente nessa perspectiva. Utilizamos a Libras como exemplo de língua de sinais para desenvolver nossa proposta.

A partir dos apontamentos feitos até aqui, esclarecemos que esse trabalho foi motivado pela seguinte questão de pesquisa: As iniciativas existentes para construção de corpora de línguas de sinais dão sustentação para que as análises que serão feitas a partir daquele conjunto de documentos se configurem de maneira eficaz? Hipótese: a falta do módulo de transcrição para articulação do sinal nas iniciativas de construção de corpora de línguas de sinais prejudica a associação, por parte do linguista e/ou de uma ferramenta computacional, entre forma e sentido, tendo como consequência prejuízo na formulação de hipóteses sobre a língua.

Ademais, as atuais iniciativas de construção de corpora para língua de sinais não reusam as tecnologias já desenvolvidas para Linguística de Corpus, o que acarreta retrabalho. A possibilidade de reuso das técnicas e tecnologias desenvolvidas pela Linguística de Corpus na construção dos corpora de línguas orais e suas duas modalidades oral e escrita dá a um corpus de língua de sinais a condição de colocá-lo em um nível mais avançado de anotação em um curto espaço de tempo. Consideramos possível o reuso de ferramentas já desenvolvidas para as línguas orais, uma vez que o que distancia os corpora de línguas de sinais dos corpora de línguas orais é a existência de um sistema de transcrição articulatória do sinal equivalente a um sistema fonético. Já temos esse sistema, desenvolvido por Lessa-de-Oliveira, o Sistema de Escrita para Línguas de Sinais – SEL, e postulamos que esse sistema, que representa a articulação do sinal de forma linear, sintética e padronizada, utilizado no módulo de transcrição do corpus é a condição que faltava para o alinhamento das anotações com o sinal. Tal alinhamento é importante para uma descrição mais precisa sobre as línguas sinalizadas, potencializando as possibilidades de explicações sobre os fenômenos linguísticos das línguas de sinais, bem como descrição de sua gramática.

Desse modo, o objetivo deste trabalho é conhecer a forma como os estudos da área da Linguística trabalham com os dados da Libras, como são constituídos os corpora que fundamentam os estudos, avaliar os limites e as possibilidades das iniciativas de construção de corpora para língua de sinais encontradas e propor um fluxo de trabalho, um workflow, para construção de corpora de línguas de sinais, que atenda todas as etapas de marcação que

segue as diretrizes dos corpora orais e escritos no que se refere ás possibilidades de anotação e que faça reuso das tecnologias já existentes. Além disso, como objetivos específicos podemos elencar: (i) verificar se a escrita Sel poderá atender aos requisitos tecnológicos da etapa de transcrição, (ii) definir os metadados que são importantes para o controle das informações para a construção do corpus, (iii) localizar as fontes de dados para a construção do corpus, (iv) propor o workflow e exemplificar sua aplicação para a construção de corpora de língua de sinais em dados de Libras extraídos de dissertações e teses consultadas na pesquisa, com o objetivo de testá-lo, (v) comparar os resultados obtidos com outras iniciativas de construção de corpora para língua de sinais e (vi) avaliar a utilização do programa ELAN[6] juntamente com a escrita SEL para o alinhamento das camadas de vídeo, transcrição e tradução na construção de corpora de Libras.

Com relação às dimensões metodológicas, essa pesquisa se enquadra numa perspectiva de Pesquisa Aplicada, tendo em vista que, como explicam Barros e Lehfeld (2000, *apud* Vilaça, 2010), esse tipo de pesquisa é influenciado pela necessidade de gerar conhecimento para aplicação de seus resultados, com a finalidade de colaborar com a solução de problemas percebidos na realidade. E temos esse intuito nesse trabalho ao propor um workflow para construção de corpora de línguas de sinais.

Sendo assim, com base nesses objetivos e nessas orientações metodológicas, executamos os seguintes procedimentos metodológicos para realização dessa pesquisa e formulação da proposta, aqui apresentados como etapas: (i) realização de um estudo para conhecimento do estado da arte e levantamento de corpora de língua de sinais observando seus alcances, possibilidades e limites; (ii) levantamento de requisitos importantes para anotação e controle na produção de corpora de língua de sinais; (iii) definição dos metadados que guiarão a catalogação dos dados, os quais serão guardados em uma estrutura de arquivo

[6] O ELAN é um software gratuito desenvolvido pelo Instituto Max Planck de Psicolinguística, projetado para a transcrição e análise de dados em áudio e vídeo. É amplamente empregado em pesquisas linguísticas, com destaque para o estudo de línguas de sinais.

como corpus cru[7]; (iv) construção de um modelo teórico a partir do workflow proposto para a construção de corpora de língua de sinais que contempla o módulo de transcrição e tradução com exemplificação da anotação a partir de dados de Libras selecionados nas fontes da pesquisa; (v) realização de análise comparativa entre o modelo construído a partir do workflow proposto e os corpora construídos com a utilização de outros esquemas para investigações da língua Libras.

A fim de apresentar a nossa pesquisa, organizamos o trabalho em seis seções: 1) Introdução – como vimos, contextualiza a pesquisa e apresenta a justificativa, os objetivos, dimensões e procedimentos metodológicos; 2) Corpora em estudos linguísticos – apresenta de maneira geral sobre corpora e sua área, e a relevância desses para estudos linguísticos; além disso, apresenta o estado da arte de corpora de Libras no que tange às suas especificidades com relação aos corpora de língua oral; 3) A escrita: representação das linguagens falada oral e sinalizada – discute um pouco sobre a história do surgimento da escrita nas línguas de cultura escrita, destaca a maneira como se dá a escrita nas línguas orais que já possuem uma escrita oficialmente estabelecida e nas línguas sinalizadas uma vez que ainda são consideradas línguas ágrafas apesar de haver iniciativas de escrita para elas; apresentamos quatro dessas iniciativas de escrita, destacando e detalhando a escrita Sel, a qual foi analisada e selecionada por nós como a escrita que melhor atende às nossas exigências para compor o módulo de transcrição; 4) Corpora eletrônicos anotados – aborda de que maneira a Linguística Computacional, a Linguística de Corpus e o Processamento de Linguagem Natural contribuem com a pesquisa linguística e demonstra a maneira como são construídos os corpora eletrônicos de línguas orais como exemplo para os corpora de Libras; 5) Apresentação da proposta de workflow para construção de corpora para línguas de sinais – ocorre a culminância da nossa pesquisa, apresentação da nossa iniciativa de construção de corpora para língua de sinais com base no Lapelinc Framework e demonstração da nossa

[7] Corpus Cru é a tradução de Corpus Raw, conceito formulado e utilizado por Santos e Namiuti (2019), entre outros trabalhos dos autores, para se referir a base de dados (fonte material do corpus) já transposta para o meio digital em formato de imagem sem edição, com anotações iniciais de catalogação, controle de procedência, estado de conservação, características físicas e tipológicas, mas ainda sem edição/anotação de edição e outras anotações que só serão aplicadas nas etapas de transcrição e compilação do corpus. Este conceito foi reutilizado e adaptado na construção do Corpus Carolina (Finger; Paixão de Sousa; Namiuti; Monte, 2021), que é um corpus aberto de textos em Português brasileiro contemporâneo (1970-2021), com informações de procedência e tipologia. O projeto Carolina faz parte do grande projeto da área de Processamento de Linguagem Natural (NLP2) do Centro de Inteligência Artificial da Universidade de São Paulo (C4IA). Ele é desenvolvido por uma equipe multidisciplinar de linguistas e cientistas da computação, membros do C4IA e do Laboratório Virtual de Humanidades Digitais (LaViHD).

iniciativa no ELAN; 6) Conclusão – discute a pertinência do trabalho, resultados e apresenta sugestões de continuação do projeto em trabalhos futuros.

2 CORPORA EM ESTUDOS LINGUÍSTICOS

Uma vez que temos por objetivo dar a conhecer o estado da arte de como são constituídos os corpora que fundamentam os estudos da Língua Brasileira de Sinais, avaliar os limites e as possibilidades dessas construções, nessa seção, nos atemos a dissertar acerca de corpora em estudos linguísticos: primeiro, dissertamos de maneira geral sobre corpora, a área da Linguística de Corpus e o quão importantes são para a pesquisa linguística; depois, tratamos sobre os corpora em estudos linguísticos da Libras, apresentando um levantamento de pesquisas em Libras, e a partir disso, discorremos acerca do processo de geração e/ou coleta dos dados e do processo de anotação dos dados.

2.1 Corpora, Linguística de Corpus e a importância dos corpora em estudos linguísticos

A palavra "Corpus", em seu sentido original, significa, de acordo com Sardinha (2004), corpo, conjunto de documentos. O pesquisador explica que os corpora tiveram dois momentos distintos com relação à sua constituição no decorrer da história: antigamente, os corpora eram construídos totalmente de maneira manual, todo processo de coleta, organização e análise era feito isento de meios eletrônicos – Alexandre o Grande, na Grécia Antiga, determinou o Corpus Helenístico; na Antiguidade e Idade Média, foram compostos corpora de citações da Bíblia. Por outro lado, atualmente, o que se vê são corpora digitais totalmente preparados com a utilização dos meios eletrônicos.

Esse avanço na tecnologia facilitou todo o manuseio com os corpora, principalmente no que diz respeito, por exemplo, à confiabilidade dos dados – uma vez que o trabalho feito por uma máquina diminui a quantidade de possíveis erros no processamento de corpora gigantescos –, ao aumento da capacidade de armazenamento, à introdução de novas mídias e à possibilidade de realizar buscas automáticas. Toda essa questão nos aponta para o fato de que antes mesmo da tecnologia, do computador, já havia a existência de corpora, entretanto, com o surgimento dessas ferramentas, o processo se tornou mais viável e eficaz.

Vale ressaltar que foi um corpus não computacional, o SEU (Survey of English Usage), compilado por Randolph Quirk e sua equipe em Londres a partir de 1959, o responsável pela formação dos corpora modernos (Sardinha, 2004). Sardinha (2004) esclarece que o fato do corpus SEU ter sido planejado para armazenar 1 milhão de palavras, ter uma definição de um número fixo de textos e também um mesmo número de palavras para cada texto, influenciou a construção de outros corpora, isto é, fez com que ele, o SEU, se tornasse

referência para outros corpora. Ainda segundo o autor, o Survey of English Usage foi organizado em fichas de papel, nas quais continha uma palavra específica em cada – cada palavra aparecia inserida em dezessete exemplos de texto. Essas palavras que compunham o corpus foram analisadas e categorizadas gramaticalmente. Inclusive, essa categorização foi utilizada como "base para o desenvolvimento dos etiquetadores[8] computadorizados contemporâneos, que fazem a identificação de traços gramaticais automaticamente" (Kader; Richter, 2013, p.4)

Mcenery e Wilson (2013) afirmam que os estudos linguísticos utilizando corpus tiveram início no final do século XIX, especificamente em 1897, apesar de alguns linguistas só terem utilizado de fato um método com base em corpus em pesquisas no século XX, nos anos 1940. Todavia, o que atualmente conhecemos como Linguística de Corpus é considerada uma expressão nova, pois segundo Leech (1992 *apud* Kader; Richter, 2013) esta foi utilizada pela primeira vez por Aarts e Meijs, em 1984, como título de um livro.

Sardinha (2004) ratifica que a Linguística de Corpus é uma área que se ocupa da coleta e exploração de corpus, de uma coleta e exploração criteriosa de dados linguísticos textuais, com o intuito de ser base para a pesquisa de uma língua ou uma variedade linguística. Desse modo, em uma pesquisa de investigação linguística, um corpus, construído sob e analisado com as orientações da Linguística de Corpus, com a grande quantidade de dados depositados nele, dá suporte ao pesquisador na verificação de padrões usados, possibilitando, assim, a descrição e explicação dos atos linguísticos, ficando a cargo do pesquisador a criação de metodologias confiáveis para analisar, descrever e explicar tais atos. (Maccarthy; O'keeffe, 2010)

2.2 Na prática: os corpora em estudos linguísticos da Libras

A fim de embasar a nossa discussão, segue, abaixo, um levantamento de pesquisas sobre Libras realizado nos repositórios digitais da Universidade de São Paulo – USP, da Universidade Estadual de Campinas – UNICAMP, da Universidade Estadual do Sudoeste da Bahia – UESB, da Universidade Federal de Rio Grande do Sul – UFRGS e da Universidade Federal de Santa Catarina – UFSC. Ao todo, foram selecionados 13 trabalhos divididos entre dissertações e teses, os quais se encontram dispostos nos quadros 1 a 13 abaixo onde destacamos características como: nome do pesquisador e do orientador, título da

[8] Ferramentas que anotam automaticamente propriedades gramaticais das palavras, como as classes gramaticais às quais pertencem.

dissertação/tese, Universidade/Programa de Pós-Graduação ao qual está vinculado, ano e local de publicação, citação direta do próprio pesquisador informando como ocorreu a geração dos dados, como se dá o acesso ao corpus e o link para ter acesso ao trabalho completo. Selecionamos esses 13 trabalhos para verificar de que forma são constituídos os corpora que fundamentam essas pesquisas.

É preciso ressaltar que essas Universidades foram designadas por conveniência por terem área para pesquisas em Libras e, nesse sentido, serem referência para o público que exerce pesquisa a respeito dessa área. Dentre tantas pesquisas de Libras encontradas nos repositórios digitais dessas Universidades, essas 13 foram selecionadas porque estudam a sintaxe e fonologia da Libras – todas que possuem esse objetivo foram, precisamente, selecionadas; pesquisas da área da educação, da área social, estudo de caso, isto é, aquelas que não trabalharam a língua em si no que se refere à sua estrutura gramatical, não foram selecionadas, pois não atendem ao nosso objetivo.

Quadro 1 – Dissertação "Uma investigação sobre o uso de recursividade em Libras"

Pesquisadora: Amanda Oliveira Rocha **Orientadora:** Ingrid Finger **Universidade/Programa de Pós-Graduação:** UFRGS/Programa de Pós-Graduação em Letras – PPGLET **Ano:** 2021 **Local:** Porto Alegre **Como os dados são gerados:** "O corpus utilizado nesta pesquisa faz parte do Inventário Nacional de Libras (Quadros2016a, 2016b, Quadros *et al*. 2017a; 2017b), que por sua vez é parte constituinte do projeto Inventário de Libras [...] Os dados que constituem o Inventário Nacional de Libras (Quadros 2016a, 2016b, Quadros *et al*. 2017a; 2017b) foram coletados com a participação de surdos brasileiros, através de vídeos-registros da Libras e, posteriormente, a equipe do projeto fez a transcrição e categorização dos sinais identificados na sinalização com ambas as mãos" (ROCHA, 2021, p. 53) **Acesso ao corpus:** Disponibilizado na rede, no link https://corpuslibras.ufsc.br/inicio **Link do trabalho:** https://lume.ufrgs.br/handle/10183/224853

Quadro 2 – Dissertação "A natureza gramatical da Libras adquirida por surdos e ouvintes: sinal, classificador, ação construída e gesto"

Pesquisadora: Thamires Oliveira de Souza Sampaio **Orientadora:** Adriana Stella Cardoso Lessa- de-Oliveira **Universidade/Programa de Pós-Graduação:** UESB/Programa de Pós-Graduação em Linguística – PPGLin **Ano:** 2020 **Local:** Vitória da Conquista

Como os dados são gerados: "[...] a coleta de dados foi realizada dividida em duas etapas: a primeira experimental, constituída por dois tipos de testes; e a segunda constituída como coleta de amostras naturalísticas através da internet. [...] Os dados experimentais foram obtidos com base na coleta de vídeos produzidos, em Libras, pelos informantes da pesquisa surdos ou por ouvintes, durante os testes de eliciações. Esses testes foram aplicados utilizando-se vídeos compostos por conteúdos distintos, sendo divididos em dois testes, assim nomeadas: Teste 1-Recorrência das ACs e CLs; e Teste2 -Grau de aceitabilidade de narrativas com sinais padrões e ACs." (SAMPAIO, 2020, p.82) **Acesso ao corpus:** Banco de dados interno **Link do trabalho:** https://repositorio.cepelin.org/index.php/repositorioppglintesesdissertaco/article/view/209/190

Quadro 3 – Tese "Iconicidade nas sentenças topicalizadas da Libras: uma motivação semântica e pragmática"

Pesquisadora: Daiana do Amaral Jeremias **Orientador:** Heronides Maurílio de Melo Moura **Universidade/Programa de Pós-Graduação:** UFSC/Programa de Pós-Graduação em Linguística da Universidade Federal de Santa Catarina **Ano:** 2020 **Local:** Florianópolis **Como os dados são gerados:** "Os dados coletados estão inseridos em uma série de vídeos catalogados pelo Núcleo de Aquisição de Língua de Sinais (NALS) da Universidade Federal de Santa Catarina (UFSC), disponibilizado publicamente no site "Corpus de Libras", disponível em: http://www.corpuslibras.ufsc.br/, mais especificamente na seção "Acervo>SC>Inventário Libras, respectivamente. [...] Buscando respostas para nossos questionamentos acerca da topicalização, selecionamos 15 vídeos de entrevistas nos quais a duração é aproximadamente entre 15 a 40 minutos. Todas as entrevistas possuem glosas dos sinais produzidos e algumas têm traduções" (JEREMIAS, 2020, p.25-26) **Acesso ao corpus:** Disponibilizado na rede, no link https://corpuslibras.ufsc.br/inicio **Link do trabalho:** https://repositorio.ufsc.br/bitstream/handle/123456789/216338/PLLG0797-T.pdf?sequence=-1&isAllowed=y

Quadro 4 – Dissertação "Aspectos de nomes e verbos na Libras: identificação morfossintática"

Pesquisador: Igor Valdeci Ramos da Silva **Orientadora:** Aline Lemos Pizzio **Universidade/Programa de Pós-Graduação:** UFSC/Programa de Pós-Graduação em Linguística da Universidade Federal de Santa Catarina **Ano:** 2020 **Local:** Florianópolis **Como os dados são gerados:** "O projeto Corpus de Libras atende aos requisitos desta investigação. Ele é uma parceria CAPES, CNPQ, IPHAN e Hiperlab, para o desenvolvimento de banco de dados que

proporcione material para pesquisa com dados e metadados da Libras “que se situa no contexto
do Inventário Nacional da Diversidade Linguística (INDL) que foi instituído pelo decreto presidencial 7387/10 como um instrumento de identificação, reconhecimento, valorização e promoção das línguas faladas no Brasil” (Quadros *et al*, 2018) e é a fonte de coleta dos dados para constituição do corpus da presente investigação.”(SILVA, 2020, p.57-58)
Acesso ao corpus: Disponibilizado na rede, no link https://corpuslibras.ufsc.br/inicio
Link do trabalho: https://repositorio.ufsc.br/handle/123456789/216534

Quadro 5 – Tese “A (in)definitude no sintagma nominal em Libras [recurso eletrônico]: uma investigação na interface sintaxe-semântica”

Pesquisador: Anderson Almeida da Silva
Orientadores: Ruth Elisabeth Vasconcellos Lopes e Josep Francisco Quer Villanueva
Universidade/Programa de Pós-Graduação: UNICAMP/Instituto de Estudos da Linguagem
Ano: 2019
Local: Campinas
Como os dados são gerados: “[...] Efetuei uma coleta de dados naturalísticos e elicitados para investigar a ocorrência, sistematicidade e interpretação de DPs nus e acompanhados de determinantes em posições argumentais. Para elicitar os dados, surdos monolíngues n=20 e surdos bilíngues n=20 participaram de três tarefas de produção e três tarefas de compreensão que tinham por objetivo identificar os itens utilizados para codificar definitude forte, definitude fraca, indefinidos específicos e indefinidos não específicos.” (SILVA, 2019, p.10)
Acesso ao corpus: Banco de dados interno
Link do trabalho:
https://www.repositorio.unicamp.br/acervo/detalhe/1126423?guid=1691493292636&returnUrl=%2fMinhaSelecao%3fguid%3d1691493292636&i=1&m=1

Quadro 6 – Dissertação “A questão da categorização morfológica para nome e verbo em Libras”

Pesquisadora: Ediélia Lavras dos Santos Santana
Orientadora: Adriana Stella Cardoso Lessa- de-Oliveira
Universidade/Programa de Pós-Graduação: UESB/Programa de Pós-Graduação em Linguística – PPGLin
Ano: 2019
Local: Vitória da Conquista
Como os dados são gerados: “[...] constituímos o corpus desta pesquisa a partir de um teste de eliciação por meio de imagens em contexto, ou seja, imagens das quais fosse possível extrair frases contextualizada [...] Após as gravações, os vídeos foram analisados e transcritos utilizando a ferramenta ELAN-EUDICO-Linguistic Annotator- 5.2” (SANTANA, 2019, p. 61, 63)
Acesso ao corpus: Banco de dados interno
Link do trabalho:
https://repositorio.cepelin.org/index.php/repositorioppglintesesdissertaco/article/view/168/146

Quadro 7 – Tese "Um Olhar da Semiótica para os Discursos em Libras: Descrição do Tempo"

Pesquisadora: Renata Lúcia Moreira **Orientadora:** Diana Luz Pessoa de Barros **Universidade/Programa de Pós-Graduação:** USP/Programa de Pós-Graduação em Semiótica e Linguística Geral **Ano:** 2016 **Local:** São Paulo **Como os dados são gerados:** "Usando diferentes estratégias para estimular as contações de história, pedimos a três diferentes colaboradores, fluentes em Libras (dois surdos e uma intérprete da língua), para criar os textos com os quais poderia dar continuidade à investigação do tempo. Selecionamos, entre os textos eliciados, cinco para compor o corpus que seria analisado com mais detalhes nesta pesquisa [...] também foram analisados e se tornaram exemplos no trabalho outros textos em Libras, eliciados durante os testes descritos na seção anterior, e aqueles que também estão disponíveis na Internet." (MOREIRA, 2016, p.98-99) **Acesso ao corpus:** Banco de dados interno **Link do trabalho:** https://teses.usp.br/teses/disponiveis/8/8139/tde13022017135649/publico/2016_RenataLuciaMoreira_VOrig.pdf

Quadro 8 – Dissertação "A interferência do Português na análise gramatical em Libras: o caso das preposições"

Pesquisadora: Myrna Salerno Monteiro **Orientador:** Tarcísio de Arantes Leite **Universidade/Programa de Pós-Graduação:** UFSC/Programa de Pós-Graduação em Linguística da Universidade Federal de Santa Catarina **Ano:** 2015 **Local:** Florianópolis **Como os dados são gerados:** "A pesquisa se alicerçou na gramática baseada no uso, com base na qual foram feitas análises de uma amostra de produções espontâneas de surdos em Libras retiradas de vídeos públicos do YouTube envolvendo o sinal PARA. Além desses dados espontâneos, foram também analisados dados do sinal PARA em compilações de sinais e dicionários de referência da Libras. Para garantir o anonimato dos autores dos vídeos, tendo em vista a dificuldade de se obter consentimento de suas produções, os vídeos coletados foram baixados, catalogados, e em seguida os trechos relevantes para a análise foram regravados pela pesquisadora, que também é surda e falante da Libras. Esses trechos analisados foram também recortados em imagens individuais dos sinais, para melhor análise e suporte à glosagem, bem como traduzidos para o Português." (MONTEIRO, 2015, p.8) **Acesso ao corpus:** Disponibilizado na rede **Link do trabalho:** https://repositorio.ufsc.br/handle/123456789/169443

Quadro 9 – Dissertação "A categoria dos verbos na língua brasileira de sinais"

Pesquisadora: Ione Barbosa de Oliveira Silva **Orientadora:** Adriana Stella Cardoso Lessa- de-Oliveira **Universidade/Programa de Pós-Graduação:** UESB/Programa de Pós-Graduação em Linguística – PPGLin **Ano:** 2015 **Local:** Vitória da Conquista **Como os dados são gerados:** "A coleta de dados foi realizada em três sessões no Centro Estadual de Educação Profissional em Gestão e Tecnologia da Informação Regis Pacheco, em Jequié. Na primeira sessão houve uma conversa informal, em que os surdos falaram um pouco de suas vidas, trabalho, estudo etc. Na segunda sessão os informantes contaram uma história infantil escolhida por eles, "Os três porquinhos". E, na terceira sessão apresentamos pares de nomes e verbos a partir de imagens para que eles sinalizassem, a fim de observarmos se havia distinção entre as categorias gramaticais ou não. Ressaltamos que todas as sessões foram individuais. Todas as sessões de coleta de dados foram filmadas [...]" (SILVA, 2015, p.20) **Acesso ao corpus:** Banco de dados interno **Link do trabalho:** https://repositorio.cepelin.org/index.php/repositorioppglintesesdissertaco/article/view/61/48

Quadro 10 – Dissertação "Sintaxe dos determinantes na língua brasileira de sinais e aspectos de sua aquisição"

Pesquisadora: Lizandra Caires do Prado **Orientadora:** Adriana Stella Cardoso Lessa- de-Oliveira **Universidade/Programa de Pós-Graduação:** UESB/Programa de Pós-Graduação em Linguística – PPGLin **Ano:** 2014 **Local:** Vitória da Conquista **Como os dados são gerados:** "De forma individual, cada informante foi convidado a narrar uma fábula, já conhecida por ele. Este procedimento foi adotado visando coletar uma amostra de fala natural, uma vez que o sujeito-informante estava livre para conduzir sua produção linguística, e livre de quaisquer interferências, seja do pesquisador, seja da busca de suporte de outra língua como referência, como o Português escrito, por exemplo, para a construção da narrativa. O objetivo foi compor o corpus deste estudo de modo que os informantes narrassem as suas histórias na estrutura e no contexto da Libras" (PRADO, 2014, p.25) **Acesso ao corpus:** Banco de dados interno **Link do trabalho:** https://repositorio.cepelin.org/index.php/repositorioppglintesesdissertaco/article/view/45/28

Quadro 11 – Dissertação "Uma descrição do processo de referenciação em narrativas contadas em língua de sinais brasileira (Libras)"

Pesquisadora: Thaís Bolgueroni Barbosa **Orientadora:** Evani de Carvalho Viotti **Universidade/Programa de Pós-Graduação:** USP/Programa de Pós-Graduação em Semiótica e Linguística Geral

Ano: 2013
Local: São Paulo
Como os dados são gerados: "A narrativa estudada nesta pesquisa, intitulada "O amor é surdo", é parte de um corpus maior de narrativas que vem sendo construído pelo grupo de estudos do gesto e de línguas sinalizadas do Laboratório de Linguagem, Interação e Cognição (LLIC) [...] A gravação da narrativa foi realizada em uma sala do prédio de Letras da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, com câmeras previamente preparadas, disponibilizadas pelo Laboratório de Apoio à Pesquisa e ao Ensino de Letras (Lapel)" (BARBOSA, 2013, p.61-62)
Acesso ao corpus: Banco de dados interno
Link do trabalho:
https://www.teses.usp.br/teses/disponiveis/8/8139/tde06052013112529/publico/2013_ThaisBolgueroniBarbosa_VCorr.pdf

Quadro 12 – Tese "Construções classificadoras e verbos de deslocamento, existência e localização na língua de sinais brasileira"

Pesquisadora: Brenda Silva Veloso
Orientador: Jairo Morais Nunes
Universidade/Programa de Pós-Graduação: UNICAMP/Instituto de Estudos da Linguagem
Ano: 2008
Local: Campinas
Como os dados são gerados: "Para proceder à análise das construções classificadoras na LSB foram utilizados dados de quatro fontes: (i) verbetes em Capovilla & Raphael (2001) denominados "classificadores"; (ii) dados reportados na literatura; (iii) dados coletados de narrativas curtas feitas por informantes surdos e (iv) dados elicitados e julgamentos de gramaticalidade de sentenças realizadas por informantes surdos." (VELOSO, 2008, p.10)
Acesso ao corpus: Disponibilizado na rede/Banco de dados interno do autor
Link do trabalho:
https://www.repositorio.unicamp.br/acervo/detalhe/436045?guid=1691493292636&returnUrl=%2fMinhaSelecao%3fguid%3d1691493292636&i=2&m=1

Quadro 13 – Dissertação "Descrição fonético-fonológica dos sinais da língua de sinais brasileira (Libras)"

Pesquisador: André Nogueira Xavier
Orientadora: Evani de Carvalho Viotti
Universidade/Programa de Pós-Graduação: USP/Programa de Pós-Graduação em Semiótica e Lingüística Geral
Ano: 2006
Local: São Paulo
Como os dados são gerados: "Para a realização dessa descrição, utilizei o dicionário de Capovilla & Raphael (2001) como minha principal fonte de dados." (XAVIER, 2006, p.76
Acesso ao corpus: Disponibilizado na rede

Link do trabalho:
https://www.teses.usp.br/teses/disponiveis/8/8139/tde18122007135347/publico/Dissertacao.pdf

Com base nesse levantamento de pesquisas que realizam um estudo gramatical sobre a sintaxe e fonologia da Libras, a seguir, discorreremos sobre a maneira como são constituídas as etapas processuais "geração e/ou coleta de dados" e "anotação" dos corpora que compõem essas pesquisas.

A respeito da primeira etapa processual mencionada, nos damos conta de que a produção de dados para a realização dessas pesquisas ocorre de duas formas distintas: (1) geração particular de dados com o uso de metodologias próprias; (2) coleta de dados em uma iniciativa de corpus maior já disponibilizado na rede.

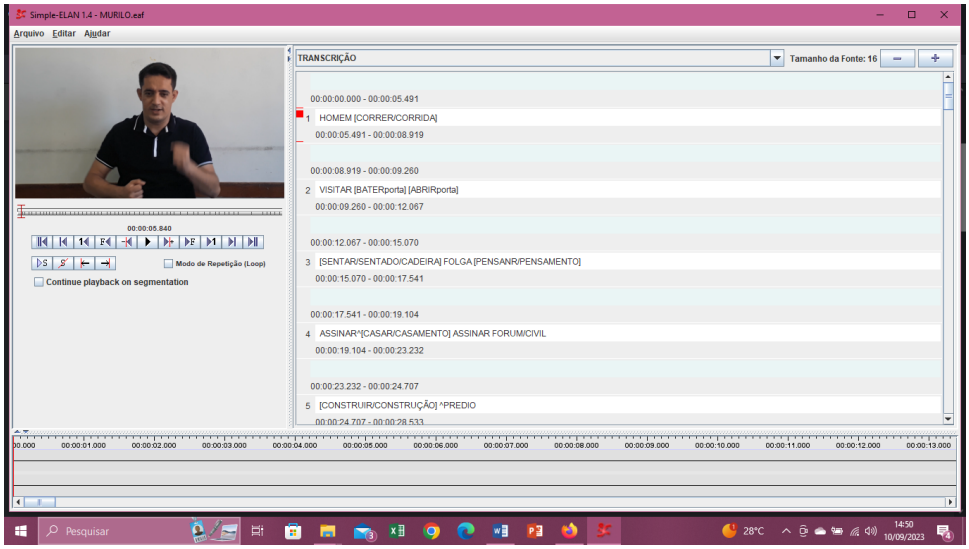
Sobre a geração de dados, essa pode ser realizada de inúmeras formas, conforme o objetivo da pesquisa. A partir da análise das pesquisas expostas acima, entra em evidência a realização de testes de eliciação de variados tipos, conversas informais, entrevistas com pergunta-resposta e gravação de narrativas. Todas essas produções são armazenadas em um banco de dados interno, banco esse que, para ter acesso, é necessário fazer contato direto para pedir permissão e o envio do material. Grande parte das pesquisas seguem esse patamar, devido à dificuldade de se encontrar iniciativas de corpora de Libras já constituídos e disponibilizados na rede.

A fim de exemplificar o que foi acima exposto, trazemos o corpus que integra uma das pesquisas mencionadas acima. A saber, esse corpus foi construído pela pesquisadora Ediélia Lavras dos Santos Santana, em 2019, na época, mestranda do Programa de Pós-Graduação em Linguística – PPGLin/UESB, para discutir em sua dissertação sobre a categorização morfológica para nome e verbo em Libras[9]. A pesquisadora explica que esse corpus foi construído "a partir de um teste de eliciação por meio de imagens em contexto, ou seja, imagens das quais fosse possível extrair frases contextualizada" (Santana, 2019, p.61). Posto isso, a pesquisadora acrescenta sobre o tratamento dos dados coletados que "Após as gravações, os vídeos foram analisados e transcritos utilizando a ferramenta ELAN-EUDICO-Linguistic Annotator- 5.2, um software de anotação que permite que se criem, editem, visualizem e procurem anotações através de dados de vídeo e áudio" (Santana, 2019, p.63)

[9] Corpus produzido de acordo com os critérios esperados por esse tipo de pesquisa, no que diz respeito às exigências do Comitê de Ética; assinatura do Termo de Consentimento Livre e Esclarecido (TCLE), número do processo 75595817.8.0000.0055.

Para ter acesso ao corpus, foi necessário, primeiramente, contatar a pesquisadora e solicitar o arquivo para download dos dados. Depois, foi preciso baixar e instalar no computador o programa ELAN. Segue a ilustração do corpus com os dados de um dos informantes:

Figura 3 – Programa ELAN, corpus de Santana (2019)



Fonte: Print Screen dos dados do informante Murilo

Cabe esclarecer que essa produção particular de dados pode envolver apenas o pesquisador, como a da pesquisadora Ediélia Lavras dos Santos Santana (2019) que acabamos de ver ou a do pesquisador Anderson Almeida da Silva (2019) que relata ter efetuado uma coleta de dados naturalísticos e elicitados para analisar a ocorrência, sistematicidade e interpretação de DPs nus que são acompanhados de determinantes em situações argumentais ou pode envolver até mesmo um grupo de pesquisadores, como o da pesquisadora Thaís Bolgueroni Barbosa (2013) que menciona que, em sua pesquisa, a narrativa estudada, “O amor é surdo”, faz parte de um corpus maior de narrativas, o qual vem sendo construído pelo grupo de estudos do gesto e de línguas sinalizadas do Laboratório de Linguagem, Interação e Cognição (LLIC). Ambos os casos se configuram como um banco de dados interno, seja do pesquisador ou do grupo. Isso acontece porque o que caracteriza esse tipo é, na verdade, a indisponibilidade de acesso para todos os pesquisadores que precisarem fazer uso do material – sendo assim restrita, ao invés de livre – e não o fato de ser uma construção individual ou em grupo.

Já com relação à segunda maneira de produção dos dados, o que ocorre, nesta, é uma coleta em uma iniciativa de corpus maior já existente e disponibilizada na rede. Aqui, todas as

pessoas têm acesso livre a esse corpus, podendo fazer uso dos dados armazenados nele conforme sua necessidade. Na tabela exposta acima verificamos a ocorrência do Corpus de Libras como exemplo disso. Destacamos que além desse, encontramos também na tabela a utilização do Dicionário de Libras de Capovilla & Raphael e vídeos públicos do YouTube para a realização das pesquisas. Entretanto, ainda que eles se encontrem disponibilizados na rede com acesso livre, esclarecemos que esses são interpretados por nós como fontes que entram em um sistema e podem ser elementos para corpus, isto é, fontes para corpus, e não um corpus constituído propriamente dito como o Corpus de Libras.

Sobre o Corpus de Libras, o encontramos disponível em: https://corpuslibras.ufsc.br/inicio. De acordo com Quadros; Schmitt; Lohn; Leite (2020), o Corpus de Libras é um conjunto de produções em Libras que resultam da união de três projetos de pesquisa: do Inventário Nacional de Libras, no qual temos diferentes apresentações do uso da Libras nos vários estados de todo o Brasil, de pessoas em três faixas etárias: de 19 até os 29 anos, 30 até 49 anos, e mais de 50 anos; de materiais de produção acadêmica produzidos por grupos de várias universidades em disciplinas dos cursos de Libras, Prolibras, disciplina de Pós Graduação, Mestrado, Doutorado, Palestras, entre outros; e das Antologias Literárias em Libras que reúnem poesias e narrativas que são contadas em língua de sinais por diferentes pessoas.

Os dados do Corpus de Libras são georeferenciados, disponibilizados por meio de vídeos e não possuem anotação. A anotação é individual, restrita e fica à cargo do projeto que constitui o corpus ou do pesquisador que deseja utilizar aquele dado; ratificamos que essas anotações não se encontram disponibilizadas no Corpus de Libras para acesso livre.

Na língua portuguesa, por exemplo, há vários corpora eletrônicos de destaque que são bem consolidados e fonte livre de pesquisa. Dentre tantos, podemos citar o Corpus Histórico do Português Tycho Brahe (disponível em: https://www.tycho.iel.unicamp.br/corpus/index.html), que é um corpus eletrônico anotado, atualmente construído com textos escritos em Português por autores nascidos entre 1380 e 1978. O Tycho Brahe conta com 95 textos (3.789.646 palavras) disponíveis para pesquisa livre, sendo que 59 desses textos possuem um sistema de anotação linguística em duas etapas, divididos da seguinte forma: 32 textos – anotação morfológica (total de 1.414.001 palavras); 27 textos – anotação sintática (total de 1.234.323 palavras).

Esse acesso irrestrito dos dados acaba enriquecendo e consolidando o próprio corpus, por meio das análises feitas das amostras retiradas do banco de dados e também da

disseminação do material de pesquisa. Além disso, evita retrabalho, uma vez que um corpus construído e de acesso livre servirá como base para inúmeras pesquisas.

Esclarecemos que, em uma pesquisa, pode ocorrer, concomitantemente, a utilização dessas duas maneiras – geração particular de dados com o uso de metodologias próprias e coleta de dados em uma iniciativa de corpus maior já disponibilizado na rede – para a construção do corpus. Exemplo disso é o caso da pesquisadora Brenda Silva Veloso (2008) que explica ter feito uso de quatro fontes para reunir dados para proceder a análise das construções classificadoras na LSB: (1) dados coletados de narrativas curtas feitas por informantes surdos e (2) outros elicitados e julgamentos de gramaticalidade de sentenças realizadas por informantes surdos – que são fontes que constituem o banco de dados interno do autor; (3) retirada de verbetes em Capovilla & Raphael (2001) denominados "classificadores" e (4) dados reportados na literatura – que são fontes de dados disponibilizadas na rede.

Tendo esclarecido a respeito da etapa processual "geração e/ou coleta de dados", agora, nos ocupamos de explanar acerca da etapa processual "anotação".

De acordo com Quadros (2016), os corpora de Libras têm em comum o registro de interações em Libras que se dá por meio de filmagens em vídeo e grande parte deles possuem também os módulos de tradução e transcrição. É preciso ressaltar que toda essa composição é crucial para que seja feita uma análise precisa e coerente dos dados: as filmagens para dar acesso à realização real gesto visual do sinal, a tradução para contribuir com a compreensão do pesquisador e a transcrição para que seja possível alcançar a associação entre forma e sentido. No entanto, essa eficácia é prejudicada nas atuais iniciativas de construção de corpora para língua de sinais pelo módulo de transcrição, uma vez que a transcrição de dados nesses referidos corpora é feita por meio de glosas de uma língua estabelecida, o que não torna possível o cumprimento da função do módulo de transcrição: fazer a associação entre forma e sentido.

Isso ocorre porque como a Libras ainda é considerada uma língua ágrafa, há uma tentativa de enquadrar a Libras em padrões de escrita de línguas orais. No entanto, essa se dá como uma tentativa falha pelo fato de estarmos diante de dados de uma língua de modalidade gesto-visual, na qual uma imagem visual toma o lugar de uma imagem acústica e a tridimensionalidade substitui a linearidade.

A fim de demonstrar essa afirmação, retomemos ao recorte do corpus construído por Santana (2019), o qual se encontra exposto acima. Como podemos observar, no presente

corpus, encontramos a filmagem do informante Murilo falando em Libras algumas frases e ao lado a transcrição por glosa dessas frases. Nesse corpus, não encontramos tradução.

Se analisarmos minuciosamente o módulo de transcrição desse corpus, percebemos que a função do módulo de transcrição não está sendo cumprida, uma vez que a transcrição por glosa não faz com que tenhamos acesso ao sentido por ela representado. Isso se verifica ao olharmos a glosagem da frase 2 dita por Murilo "VISITAR [BATERporta] [ABRIRporta]"; ela não nos fornece indícios de como realizar essa frase em Libras tal como ela é, ou seja, olhando para a forma não conseguimos acessar o sentido. Em outras palavras, se percebe que falta, nas iniciativas de construção de corpora de língua de sinais, o módulo de transcrição, prejudicando assim a associação, por parte do linguista e/ou de uma ferramenta computacional, entre forma e sentido, tendo como consequência prejuízo na formulação de hipóteses sobre a língua. Isso porque, o que se encontra nessas iniciativas é, na verdade, um módulo que exerce uma função mais parecida com a do módulo de tradução, uma vez que, devido à falta de uma escrita para as línguas de sinais efetivamente em uso, o que acontece é que os pesquisadores acabam utilizando uma língua estabelecida.

Para sintetizar, Johnston (1991) explica que o uso da glosa acarreta em desvantagens: "a relação idiossincrática entre as glosas e a produção dos sinais, a não captação da realização do sinal físico, a necessidade de descrição da glosa, bem como o fato de a glosa poder indicar coisas que não estão representadas no sinal e poder ser insuficiente para representar o sinal" (Johnston, 1991 *apud* Quadros, 2016, p. 15)

Nos corpora de língua portuguesa em que o objeto de investigação é a oralidade, tal problema não é recorrente no módulo de transcrição, tendo em vista que essa não é língua ágrafa, já existe um sistema de escrita que dá conta de escrever e de transcrever a língua portuguesa. Nesses corpora encontramos uma gravação da fala dos informantes – nesse caso não precisa necessariamente ser por vídeo, pois esta é uma língua oral, e não gesto visual como a Libras – e uma transcrição fonética desses dados. Por exemplo, se o informante fala na gravação a palavra "rasgado", essa palavra será representada no módulo de transcrição por uma das quatro seguintes possíveis formas fonéticas: (1) com a fricativa velar sonora [ɣ] e a fricativa alveolar sonora [z] = [ɣaz'gadʊ]; (2) com a fricativa velar sonora [ɣ] e a fricativa pós-alveolar sonora [ʒ] = [ɣaʒ'gadʊ]; (3) com a fricativa velar surda [x] e a fricativa alveolar sonora [z] = [xaz'gadʊ] e (4) com a fricativa velar surda [x] e a fricativa pós-alveolar sonora [ʒ] = [xaʒ'gadʊ]. Quem conhece os comandos da escrita fonética, é capaz de olhar para a transcrição e realizar a palavra tal como ela é realizada na gravação, com todos os sons, especificidades e variações dialetais, sem fazer inferências. Nesse exemplo da palavra

“rasgado”, a escolha por utilizar uma dessas quatro possibilidades fica à cargo da pronuncia do falante.

Desse modo, nesse tipo de corpus, a função do módulo de transcrição é cumprida com êxito; em outras palavras, a transcrição fonética nos corpora de língua oral em que o objeto de investigação é a oralidade dá conta de explicar precisamente a forma como se produz um som linguístico, uma palavra, algo que a glosa nos corpora de Libras não é capaz de alcançar.

Frente a esses apontamentos, é nítido que se pensar em um workflow para construção de corpora em línguas de sinais é fundamental. Nessas circunstâncias, esclarecemos que o workflow que se objetiva alcançar como resultado desta pesquisa servirá como orientação metodológica para as iniciativas de construção de corpora em línguas de sinais, tanto para as construções individuais restritas como a da pesquisadora Ediélia, quanto para um projeto que associe essas pesquisas pragmáticas na construção de corpora abertos que atenda a comunidade e a pesquisa em sentido mais amplo, a exemplo do Corpus de Libras. Esse workflow é uma proposta que ajudará a padronizar a atividade de construção de corpora para línguas sinais e, assim, solucionar problemas como a ausência ou incoerência de um módulo de anotação.

3 A ESCRITA: REPRESENTAÇÃO DAS LINGUAGENS FALADAS ORAL E SINALIZADA

> [...] para que surdos sejam bons leitores da língua oficial de seu país, importantíssimo é que sejam ótimos leitores em sua própria língua [...]. (Silva, 2012, p. 201).

Na presente seção, trazemos, primeiramente, uma discussão sobre o surgimento da escrita das línguas de cultura escrita, destacando a importância dessa para as línguas de modo geral e, a partir disso, compreendemos o quão urgente é que se consolide um sistema de escrita para as línguas de sinais, contribuindo para que elas se constituam, cada vez mais, como, de fato, línguas, no que diz respeito, por exemplo, à cultura e à produção e divulgação do conhecimento. Além disso, explanamos acerca da diferença existente entre as construções de escrita das línguas de cultura escrita e propostas de construções de escrita para as línguas de sinais. Depois, discutimos que, embora seja uma questão urgente essa necessidade de se estabelecer uma escrita para as línguas de sinais, a escrita é um desafio para os surdos e para essas línguas; mas iniciativas de sistemas de escrita já surgiram e vem sendo testadas. Sendo assim, posteriormente, apresentamos e discorremos sobre as quatro iniciativas de sistemas de escrita de sinais presentes no Brasil. Em seguida, explicitamos o motivo pelo qual selecionamos, dentre essas quatro iniciativas de sistemas de escrita, a escrita Sel para compor a etapa de transcrição na nossa proposta de fluxo de trabalho para construção de corpora para língua de sinais. Para finalizar, fazemos um detalhamento da escrita Sel, no que diz respeito às suas regras e caracteres.

3.1 A escrita nas línguas de cultura escrita

De acordo com Sampaio (2009), foi por volta de 3500 a.C. que surgiram os chamados embriões da escrita, quando os humanos começaram a gravar sinais e figuras em superfícies como pedras, lajes, paredões de falésias e paredes de caverna. Segundo o autor, isso aconteceu devido à necessidade inerente e incontrolável que os seres humanos têm de se comunicar e de se expressar. Esses sinais e figuras gravados são nomeados de arte rupestre, e, inclusive, dentre outros, “A gruta de Lascaux, na Fraça, e a de Altamira, na Espanha, entalhadas por volta de 30000 a.C., abrigam muitos dos mais antigos e ricos tesouros da arte rupestre” (Sampaio, 2009, p.32). No Brasil, que teve ocorrências da arte rupestre aproximadamente 35000 anos a.C., ganha destaque a da Serra da Capivara, no Piauí. Em

consonância com Sampaio (2009), “a arte rupestre é a mais antiga expressão da Humanidade” (Sampaio, 2009, p.32).

Essa arte rupestre se caracteriza como uma escrita pictográfica e é considerada como o primeiro tipo de comunicação gráfica. Nesse tipo de escrita, de acordo com Sampaio (2009), cada símbolo, expresso por desenho ou diagrama, representa um conceito ou palavra. Uma dificuldade encontrada nessa escrita pictográfica é que palavras que exprimem ideias, que não possuem um significado concreto, são incapazes de serem escritas. Nesse sentido, surgem dela, então, a escrita ideográfica ou analítica.

Na escrita ideográfica, cada sinal, seja ele figurativo ou geométrico, representa a notação de uma palavra. Essa é, na verdade, um aperfeiçoamento da pictográfica, uma vez que “todos os sistemas ideográficos derivam dos sistemas de escrita pictográfica, com representações estilizadas de conceitos abstratos sendo acrescentados à lista de símbolos” (Sampaio, 2009, p.34). A primeira escrita desse tipo foi a Cuneiforme, criada na Mesopotâmia pelo povo Sumério e adotada pelos acádios, assírios, medos, persas e outros. Com o passar do tempo, a escrita cuneiforme foi sofrendo modificações no que diz respeito à sua maneira de representar as ideias e, também, ao surgimento de uma nova forma silábica da escrita, na qual símbolos eram agrupados para constituir nomes próprios em razão do som semelhante comum entre eles.

Depois dessa, surge a escrita consonântica, que é caracterizada como um sistema alfabético-fonético, na qual, como é explicitado pelo próprio nome, consta somente consoantes. Temos como exemplo de escrita consonântica a escrita fenícia que possuía de vinte a trinta letras. Nessa, de acordo com Sampaio (2009), os sinais indicam um elemento totalmente sonoro, letras com traços relativamente simples, perdendo, assim, a configuração concreta dos símbolos. Mais tarde, essa escrita consonântica se aperfeiçoou, dando lugar ao surgimento da escrita alfabética que contava com consoantes e também vogais, utilizada, por exemplo, pelos gregos, latinos e etruscos.

Esse tipo de sistema de escrita que tem o som da fala como base começou a ser tido como uma forma eficiente de representar a simbologia intrínseca nas palavras, em detrimento de sistemas de escritas fundamentados no conceito.

Moreira e Rosado (2020) explicam que antes de existirem os sistemas de registro escrito para as línguas orais, as comunicações eram feitas por meio de mensageiros que as repassavam oralmente para um grupo ou indivíduo. Todavia, com o surgimento dos sistemas de escrita para essas línguas, as mensagens tiveram condições de alcançar, em menos tempo, um número muito maior de pessoas; na época, através de tabuletas, pergaminhos, papiros ou

livros. Atualmente, a troca de informações acontece de forma quase instantânea com a utilização de meios eletrônicos como computadores e celulares. Porém, enfatizamos que isso ocorreu e ocorre com as línguas orais. Quando nos referimos às línguas de sinais, a realidade é outra, uma vez que essas ainda são consideradas línguas ágrafas.

A escrita de uma língua é crucial em sentidos que envolvem, por exemplo, o papel cultural, político, social e educacional. Entretanto, como pontua Moreira e Rosado (2020), além disso, a escrita de uma língua funciona também como "um instrumento de registro de memória (cognição) que pode marcar todo o processo de construção da língua que representa graficamente, levando quem escreve a pensar e repensar suas estruturas formais" (Moreira; Rosado, 2020, p.198). Daí, então, vemos a urgente necessidade das línguas de sinais perderem esse status de línguas ágrafas, passando a se constituírem como línguas providas de escrita.

3.2 A escrita como um desafio para os surdos e para as línguas de sinais

A língua brasileira de sinais é uma língua que foi reconhecida oficialmente como, de fato, um meio legal de comunicação e expressão dos surdos, muito recentemente, especificamente em 24 de abril de 2002, pela Lei nº 10.436. Por essa razão, essa língua ainda carece de estudos em todos os seus âmbitos, inclusive no que tange à consolidação de uma modalidade escrita para ela.

Por a Libras ainda não dispor de uma modalidade escrita consolidada, isto é, ser até então uma língua ágrafa, os surdos precisam se valer de uma Segunda Língua (L2), como o Português, em ambientes de escrita. Essa questão se torna uma dificuldade para os surdos, uma vez que, para eles, a aquisição da modalidade escrita significa, na verdade, a alfabetização em uma outra língua que tem mecanismos de funcionamento linguístico muito diferentes, tendo em vista que a Libras é uma língua gesto-visual e as línguas orais tem uma escrita alfabética muito ligada à articulação oral. Em outras palavras: para o surdo, a aquisição da escrita não significa somente a aquisição de mais uma modalidade da língua como ocorre com os ouvintes, pois, diferentemente do ouvinte, o surdo não possui a condição auditiva para fazer a associação entre os fonemas e os grafemas, algo que é muito utilizado pelos ouvintes na aquisição da escrita, apesar das propriedades pertencentes a cada modalidade.

Como resultado dessa dificuldade, o que se encontra, de acordo com Moreira e Rosado (2020), é uma escrita do Português pelo surdo muito distante do padrão da norma culta, ainda que este esteja inserido em uma escola ideal e bilíngue.

Portanto, entende-se, com base nesse contexto, que os surdos são obrigados a serem bilíngues, pois utilizam as línguas de sinais, sua Primeira Língua (L1), para falar e precisam utilizar os mecanismos de uma língua oral, como L2, em situações nas quais a escrita é requisitada. Frente a esses apontamentos, entendemos que há uma urgência em se consolidar uma escrita que seja eficaz para a Libras, para que esse caráter bilíngue do surdo não seja algo obrigatório, mas sim alternativo.

Entretanto, como afirma Moreira e Rosado (2020):

> O registro escrito de uma língua relativamente desconhecida, como é o caso da Língua Brasileira de Sinais (Libras), dentro de um país como o nosso, o Brasil, que possui uma língua oficial (a Língua Portuguesa) utilizada pela maioria da população, encontra muitas barreiras e conflitos. São conflitos tanto na ordem jurídica (reconhecimento de língua minoritária) quanto no contexto educacional (estudo e ensino dessa língua), até aqueles relativos às suas especificidades linguísticas e status social adquirido (Moreira; Rosado, 2020, p.188).

Enfrentando as várias barreiras e conflitos existentes com relação ao registro escrito de uma língua relativamente desconhecida e na intenção de solucionar os problemas oriundos deles, surgiram, assim, as iniciativas de sistemas de escrita para a Libras.

3.3 Iniciativas de sistemas de escrita de sinais para a Libras

No Brasil, em consonância com Silva *et al*. (2018), existem quatro iniciativas de sistemas de escrita de sinais sendo experimentadas. A saber, essas são: o sistema SignWriting (SW) de origem estadunidense; e a Escrita de Língua de Sinais (ELiS), a Escrita Visogramada das Língua de Sinais (VisoGrafia) e o Sistema de Escrita da Libras (Sel) de origem brasileira. Vale a pena retomar que, como já foi mencionado em nota, essas são iniciativas de escrita para as línguas de sinais usadas por enquanto apenas na academia para fins de pesquisa. Nenhum desses sistemas está ainda efetivamente em uso pela comunidade surda/falante da língua; elas se configuram como possíveis escritas.

3.3.1 O sistema SignWriting (SW)

O SignWriting, de acordo com Moreira e Rosado (2020), foi criado em 1974 pela coreógrafa norte-americana Valerie Sutton. Ele foi introduzido no Brasil em 1996 e é o sistema mais conhecido no momento. Segundo os autores, para formar as escritas dos sinais

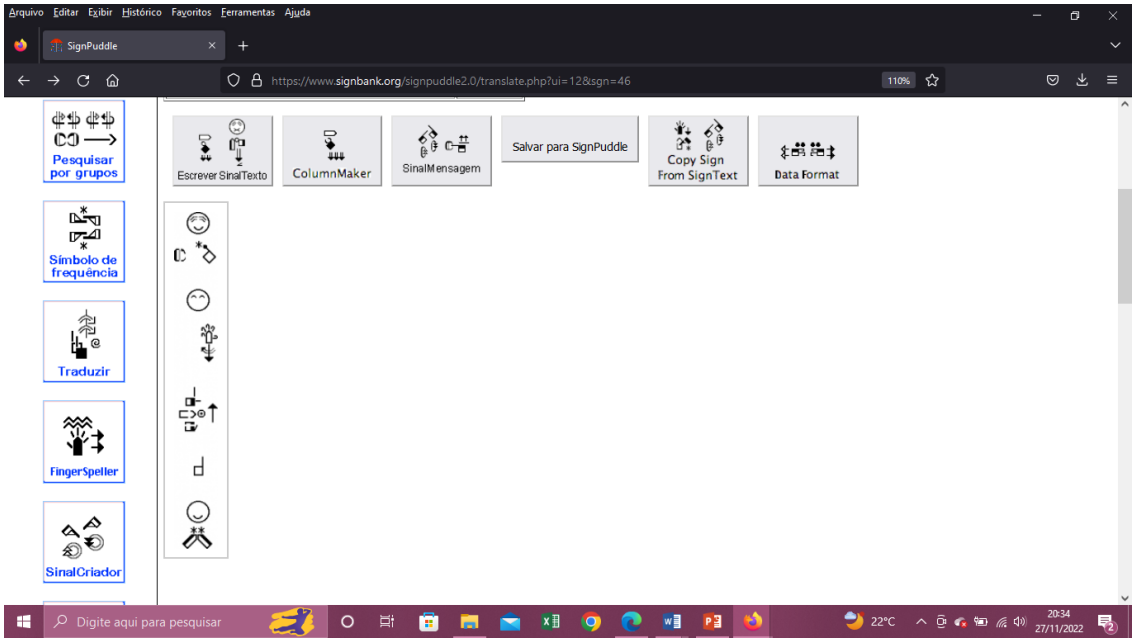
nesse sistema, faz-se necessário um software, sendo o "SignPuddle Online" – o qual acessamos em https://www.signbank.org/signpuddle2.0/signmaker.php?ui=12&sgn=46 –, o mais utilizado, atualmente. Aqui, os sinais podem ser escritos tanto de cima para baixo, quanto da esquerda para a direita.

Conforme Silva *et al*. (2018), o SW se define como um sistema gráfico-esquemático-visual e secundário das línguas de sinais, no qual se adota, em sua grande maioria, grafemas iconográficos, com aspectos gráficos e esquemáticos analógicos. Nesse sentido, os estudiosos esclarecem que:

> O SW é dividido em dez categorias: mãos, contato das mãos, faces, movimentos do corpo e da cabeça, ombro, membros, inclinação da cabeça, localização, movimento de dinâmicas e pontuação. Estas categorias são divididas em grupos [...]
> A estrutura é composta de informações referentes às mãos, movimento, expressão facial e corpo. As informações das mãos, direita e esquerda, consistem em configuração da mão, dos dedos e do braço. O movimento pode ser dos dedos (movimento interno) ou da mão (movimento externo). Um movimento pode ser composto de um ou mais movimentos de dedos, movimentos de mãos e contatos.
> A estrutura contém informações sobre a expressão facial, formada por expressões e movimento das diversas partes do rosto [...]
> Quanto às partes do corpo, a estrutura é formada por informações referentes às configurações e movimentos do ombro, tronco e cabeça. Três configurações básicas de mão: mãos circular (punho aberto), aberta (mão plana) e fechada (punho fechado). O SW tem sete símbolos que podem representar a mão sem especificar se essa mão é a direita ou a esquerda. Existem seis formas de representar o contato dos símbolos que compõe o sinal, seja mão com mão, mão com corpo ou mão com cabeça (Silva *et al*., 2018, p. 5-7).

Como forma de exemplificar as explanações até aqui feitas sobre esse sistema, segue a escrita em SW da frase "Eu quero comprar uma casa", traduzida no software "SignPuddle Online".

Figura 4 – Escrita SW da frase “Eu quero comprar uma casa



Fonte: Print Screen da página do software “SignPuddle Online”

Com base na figura acima, podemos observar, como primeira característica da escrita do SW, que os sinais se organizam na frase em sequência linear, de cima para baixo, o que, como vimos é uma das duas possibilidades de orientação nesse sistema. Ou seja, os sinais são escritos de forma linear, em colunas. A linearidade nesse sistema fica, todavia, no nível das frases, isto é, não se estende à estrutura interna da escrita dos sinais, fincando os símbolos gráficos, nesse nível, dispostos no plano bidimensional à moda de um desenho, um quadro, uma pintura.

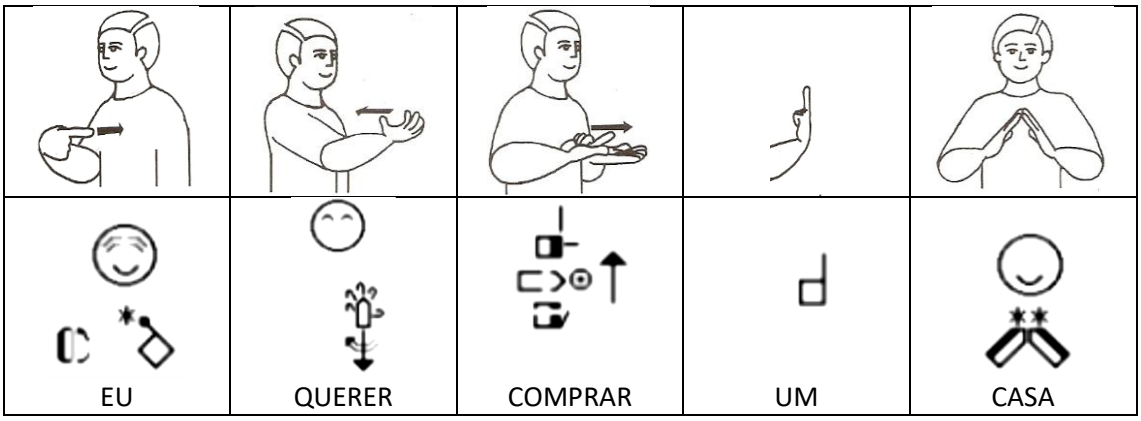
Outra observação que fazemos é que, apesar de os caracteres da SW representarem segmentos articulatórios (como configuração de mão: , , , , , , tipos de toque /movimento/direção: , ,), não observamos uma correlação muito fiel aos segmentos articulatórios constitutivos dos sinais como, por exemplo, nos indica a presença dos caracteres , , , respectivamente nos sinais EU, QUER, e CASA. Esse caracteres, que iconicamente nos remetem à ideia de uma cabeça/rosto com uma expressão facial, não correspondem a segmentos desses sinais, os quais não envolvem a cabeça, o rosto nem uma expressão facial em suas constituições articulatórias, e

sem uma causa clara, nos sinais  COMPRAR e  UM, a dita ‘cabeça’ não aparece. Em contrapartida, o sinal EU apresenta um segmento – o *tórax* que é tocado pela ponta do dedo indicador – o qual não está representado por nenhum caractere nessa escrita. Podemos tentar considerar que o tórax seria esse “vazio” abaixo do caractere que representa a ‘cabeça’. Mas o que impediria uma decodificação equivocada do sinal CASA, interpretando-se a ocorrência de um toque das duas mãos no tórax, indicado pelo símbolo  (que significa toque), assim como em EU esse símbolo () está indicando toque de uma das mãos no tórax?

O que parece haver é uma forte concentração da composição do sistema na iconicidade. Podemos perceber isso com mais clareza ao correlacionar a escrita dos sinais da frase “Eu quero comprar uma casa” com figuras da realização desses sinais, como na figura a seguir.

Assim, encontramos, na SW, uma combinação de caracteres que tem a finalidade de representar a articulação de um sinal muito parecida com uma tentativa de desenhar a fala sinalizada.

Figura 5 – Percepção icônica da SW



Fonte: Elaboração própria com imagens de Capovilla *et al.* (2011)

Consoante Silva *et al.* (2018), no sistema SW, como em qualquer outro sistema de escrita, ocorre modificações e aperfeiçoamentos. Isso acontece, segundo os autores, com sugestões importantes de pessoas surdas que têm as línguas de sinais como Primeira Língua (L1) e de membros da comunidade surda, com base em análises, estudos e na própria utilização do sistema.

3.3.2 A Escrita de Língua de Sinais (ELiS)

A ELiS é um sistema de escrita para língua de sinais, proposto por Barros (2008), que se configura como base alfabética linear, da esquerda para a direita, organizada com base nos parâmetros dos sinais propostos por Stokoe em 1965. Nesse sistema, a escrita dos elementos das línguas de sinais recebe uma nomenclatura específica: visografemas, que significa unidades mínimas (-ema) escritas (graf-) dos visemas (vis-).

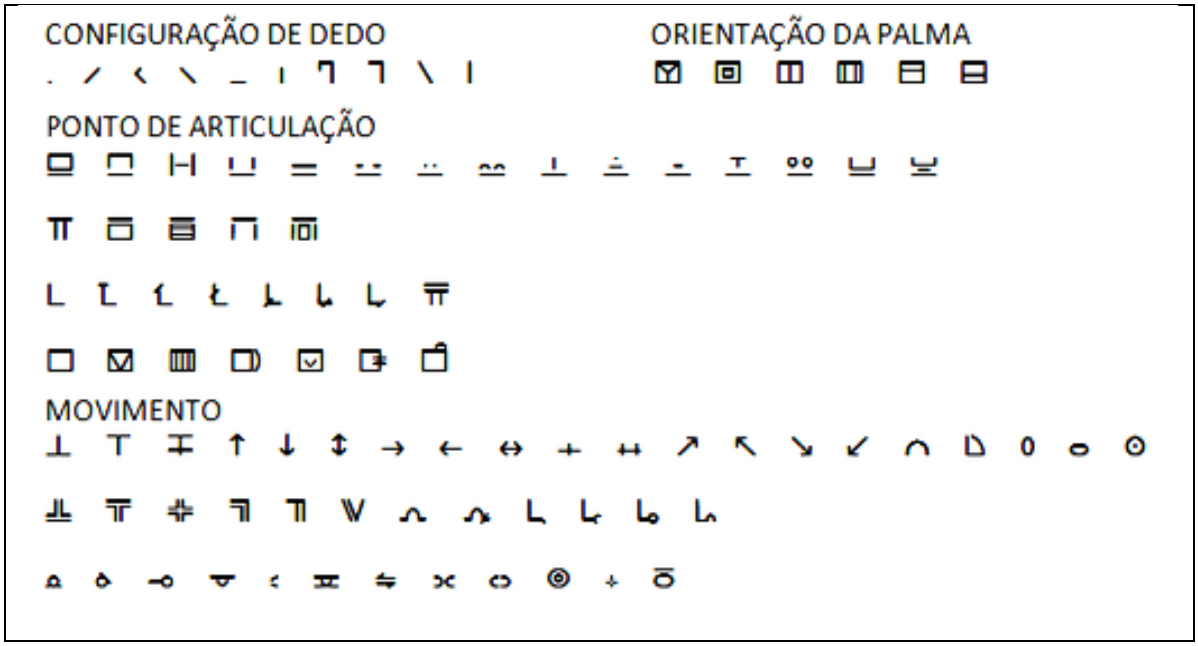
Apesar de ter se baseado nos parâmetros dos sinais propostos por Stokoe, a ELiS se distingue desses, principalmente, no que diz respeito à substituição do parâmetro Configuração de Mãos pelo parâmetro Configuração de Dedos. Sendo assim, de acordo com Barros (2008), na ELiS, a escrita é formada por quatro parâmetros: Configuração de Dedos (CD), Orientação da Palma (OP), Ponto de Articulação (PA) e Movimento (Mov).

Ainda segundo a autora, na ELiS, há 90 visografemas e eles são agrupados da seguinte maneira:

> 10 visografemas no parâmetro CD, sendo 5 para representações do polegar, 4 para os demais dedos, e 1 em comum. 6 visografemas no parâmetro OP. 35 visografemas no parâmetro PA, sendo 16 para representações de PA da cabeça, 6 do tronco, 6 dos membros, e 7 separadamente para a mão. 39 visografemas no parâmetro Mov, sendo 17 para movimentos externos da mão, 11 para movimentos internos da mão e 11 para movimentos realizados sem as mãos (Barros, 2008, p. 28-29).

Esse quadro de visogramas foi ampliado para 95, conforme o quadro apresentado por Barros (2016), exposto na figura a seguir. Quando vamos escrever uma palavra usando a ELiS, escrevemos primeiro os visografemas de Configuração de Dedos (CD), em seguida a Orientação da Palma (OP), Ponto de Articulação (PA) e por último o Movimento (M). Para os sinais que não apresentam movimento, não é necessário a escrita do último grupo.

Quadro 14 – Caracteres da Elis

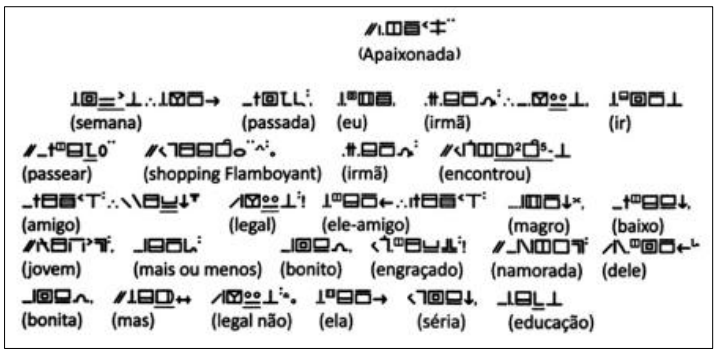


Fonte: https://shortlurl.com/Xu1w

Sobre a viabilização da escrita digital em ELiS, Barros (2008) explica que é necessário baixar a fonte True Type, que foi desenvolvida por Peixoto (2007), no arquivo de fontes de um computador e seguir a orientação de correspondências das teclas. Assim, é possível se fazer uma digitação com um teclado comum, sem precisar de programas específicos para sua utilização.

Segue, abaixo, um exemplo de escrita em ELiS retirado do texto de Barros (2008).

Figura 6 – Exemplo de escrita ELiS



Fonte: Print Screen de Barros (2008)

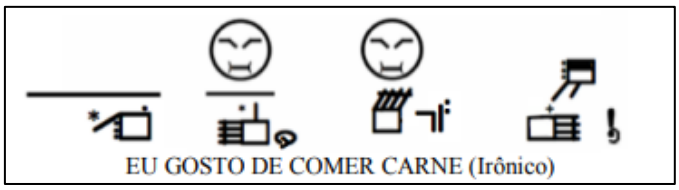
3.3.3 A Escrita Visogramada das Língua de Sinais (VisoGrafia)

A VisoGrafia é um sistema de escrita para língua de sinais que foi desenvolvido recentemente, em 2016, pelo pesquisador Claudio Alves Benassi. Consoante Silva *et al.*

(2018), o sistema da VisoGrafia une os elementos simples e visuais do SW com uma simplificação de elementos presentes na ELiS. Esse sistema grafa a língua linear e sequencialmente da esquerda para a direita, seguindo a ordenação dos visemas, ou também chamados de parâmetros: Configuração de Mão (CM), Locação (L), Movimentos (M), Direção ou Orientação da Palma (OP) e as Expressões Não Manuais (ENM).

Segue, abaixo, a frase “Eu gosto de comer carne” em escrita VisoGrafia, retirada de Benassi (2018).

Figura 7 – Escrita VisoGrafia da frase “Eu gosto de comer carne”



Fonte: Print Screen de Benassi (2018)

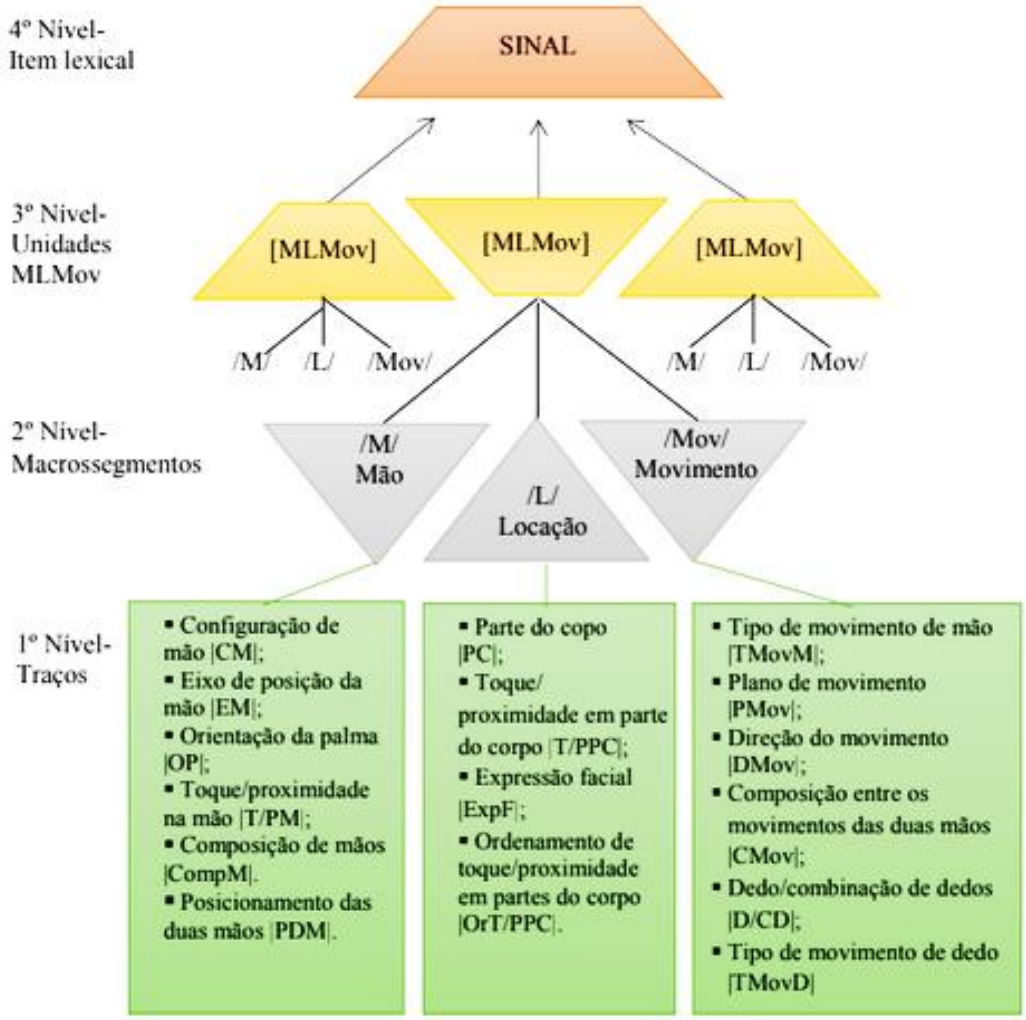
De acordo com Benassi (2018), o número de visografemas atual é 37 e para realizar uma escrita que esteja em um estágio mais básico é necessário utilizar somente esses visografemas combinados com diacríticos de configuração de dedo, contato e movimento, o restante é dispensável, contribuindo, assim, para uma escrita mais bem elaborada.

3.3.4 O Sistema de Escrita da Libras (Sel)

O Sistema de Escrita da Libras (Sel) começou a ser desenvolvido em 2009 pela linguística Adriana Stella Cardoso Lessa-de-Oliveira, obtendo uma primeira versão em 2011. Segundo Lessa-de-Oliveira (2023), essa escrita é composta por segmentos articulatórios a partir de letras e diacríticos, que são ordenados obrigatoriamente de forma linear da esquerda para a direita.

Para desenvolver a escrita do sinal em Sel, sua autora realizou paralelamente, conforme explica Lessa-de-Oliveira (2023) na apresentação da obra, uma investigação sobre a estrutura articulatória do sinal, chegando à proposta de um modelo fonológico, de acordo com o que o sinal da Libras é constituído articulatoriamente em quatro níveis: 1º dos Traços, 2º dos Macrossegmentos, 3º das Unidades MLMov e 4º do Item lexical, conforme o diagrama abaixo:

Figura 8 – Estrutura fonológica do sinal



Fonte: Lessa-de-Oliveira (2023, p.70)

Assim, com base nesse modelo, os traços do primeiro nível compõem três tipos de macrossegmentos no segundo nível: Mão /M/, Locação /L/ e Movimento /Mov/. Esses macro segmentos formam as unidades MLMov, que são, nessa perspectiva, as unidades articulatórias básicas das línguas de sinais e trazem na sua estrutura não mais que um desses macrossegmentos podendo ocorrer, em Libras, nos seguintes arranjos: [MLMov], [ML], [MMov] ou [M]. Essas unidades dão conta de marcar cada traço da configuração tridimensional do sinal. De acordo com Lessa-de-Oliveira (2023), o macrossegmento Mão é composto pelos traços de configuração de mão |CM|, eixo de posição da mão |EM|, orientação da palma |OP|, toque/proximidade na mão |T/PM|, composição de mãos |CompM| e posicionamento das duas mãos |PDM|. Já o Locação é composto por parte do corpo |PC|, toque/proximidade em parte do corpo |T/PPC|, expressão facial |ExpF| e ordenamento de toque/proximidade em partes do corpo |OrT/PPC|. Por fim, o macrossegmento Movimento é

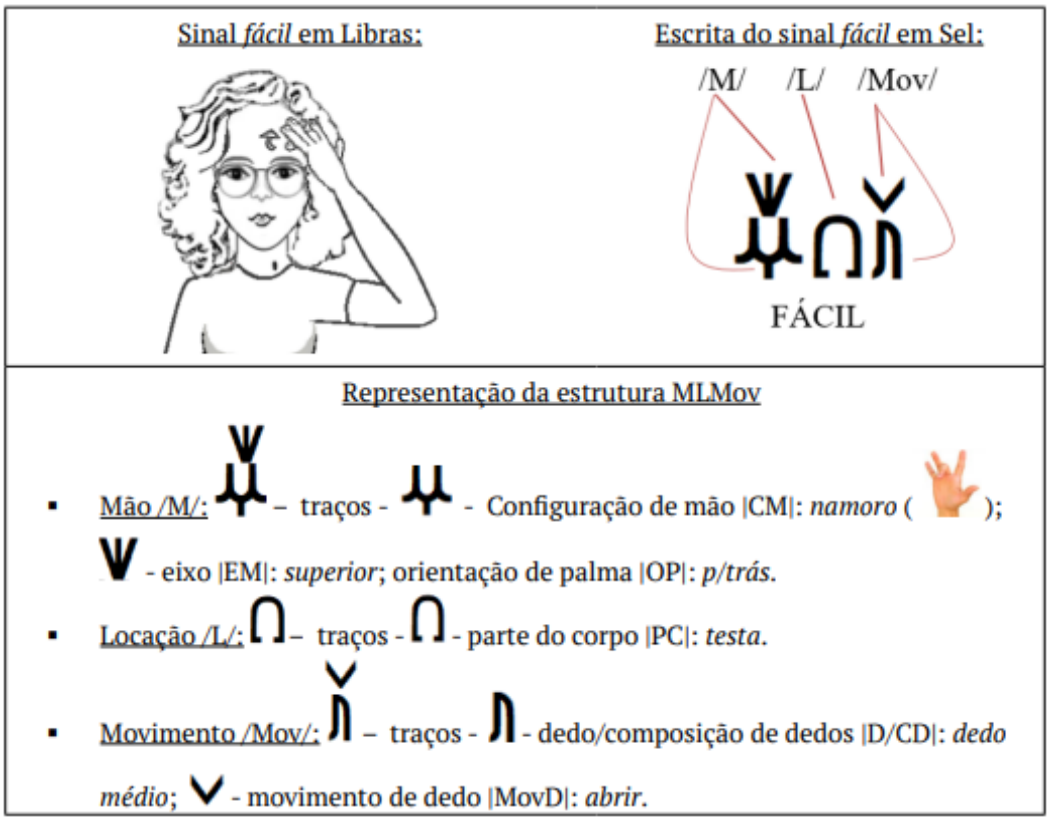
composto pelos traços de tipo de movimento de mão |TMovM|, plano de movimento |PMov|, direção do movimento |DMov|, composição entre os movimentos das duas mãos |CMov|, dedo/combinação de dedos |D/CD| e tipo de movimento de dedo |TMovD|. Conforme a autora, grande parte dos sinais em Libras é formada por apenas uma unidade MLMov, porém encontramos também sinais com duas e mais raramente com três dessas unidades. Ela não verificou nem sinal com mais que três unidades MLMov, e afirma que é muito pouco provável que ocorram sinais constituídos com mais que três dessas unidades em Libras.

A escrita Sel foi produzida tomando-se como base esse modelo fonológico do sinal. Assim, constituindo-se como um sistema de escrita trácico-fonêmica, conforme Lessa-de-Oliveira (2023, 99-100):

> a escrita Sel representa o sinal através de caracteres que correspondem aos elementos do 1º nível da estrutura fonológica MLMov. Isto quer dizer que os caracteres (letras e diacríticos) da Sel representam diretamente os traços distintivos que formam o sinal. Sendo assim, trata-se de um sistema trácico. Mas, pela forma como esses caracteres se organizam, distribuídos como letras e diacríticos fixados nos espaços superior e inferior das letras, enxergamos nesses conjuntos – de letras + seus diacríticos correlacionados – os macrossegmentos, que pertencem ao 2º nível fonológico do sinal. Vemos representação do 2º nível também em alguns casos como os de letras que indicam movimentos em planos, em que uma letra sozinha reúne três traços (tipo de movimento de mão |TMovM|, plano de movimento |PMov| e direção do movimento |DMov|), ainda que possamos distingui-los claramente em pedacinhos dessas letras.

Assim, como regra geral, a escrita Sel representa os segmentos fonológicos dos 1º e 2º níveis da estrutura fonológica MLMov por meio de letras e diacríticos, que dispõem os macrossegmentos obrigatoriamente na seguinte ordem da esquerda para a direita: 1º /M/, 2º /L/ e 3º / Mov/, como se verifica na figura a seguir.

Figura 9 – Realização do sinal "fácil" e sua escrita em Sel



Fonte: Adaptado de Lessa-de-Oliveira (2023, p. 101)

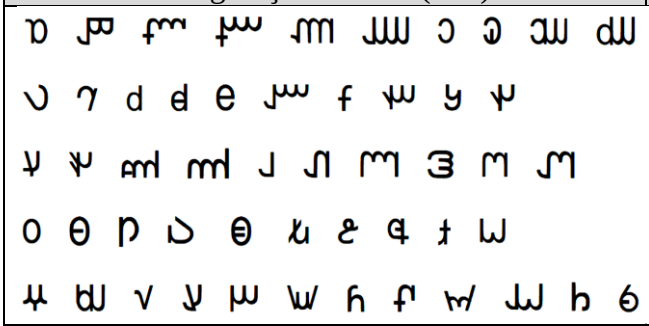
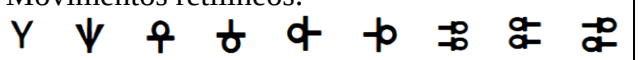
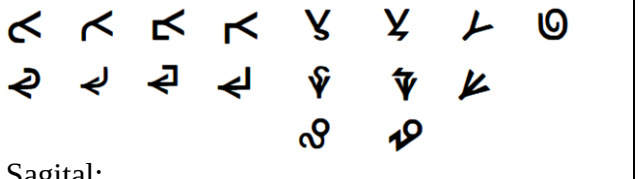
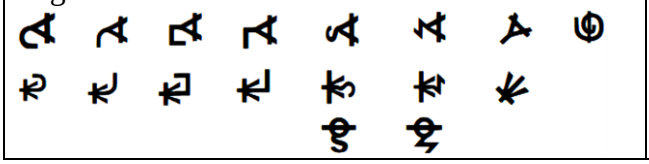
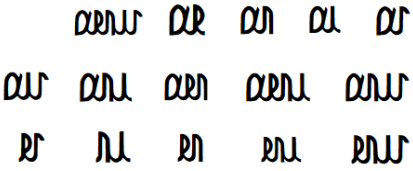
Como se pode observar nessa figura, os traços de cada macrossegmento se dispõem conjuntamente em letras e diacríticos correlacionados a essas letras. Isso e a ordem fixa na posição de representação de cada macrossegmento procuram contribuir para o automatismo do processamento na leitura e na escrita, conforme a autora.

O conjunto de caracteres do sistema Sel distribui-se em 133 letras[10] e 70 diacríticos, expostos no quadro abaixo. Esses caracteres podem representar um único traço da articulação do sinal, como as letras de *configuração de mão* e *partes do corpo* e os diacrítico de *posição das duas mãos, pontos de toque/proximidade, movimento/toque alternado ou conjunto, movimento retilíneo brevíssimo e contido,* e *expressões não manuais*, ou podem representar mais de um traço, como as letras de *movimento de mão*, cuja cauda da letra representa o traço tipo de movimento e o formato da cabeça da letra sozinho ou combinado com a posição da letra representa os traços *plano* e *direção do movimento* (ou só *direção*, no caso dos movimentos retilíneos), fazendo exceção ao movimento *circular*, que tem esses três traços representados de outra maneira, e aos movimentos fora de planos, que representam apenas o

[10] Nessa contagem entram apenas as diferentes formas dos caracteres, não entram as repetições dos caracteres de mesma forma em posições invertidas nem as combinações dos caracteres de dedos.

traço tipo de movimento. Além dessas duas formas de representação de traços, ocorre também entre os caracteres da Sel a representação de um traço por mais de um caractere. É o que ocorre na representação de movimento de dedos em que tal traço é representado pela letra de dedo(s) envolvido(s) no movimento e um diacrítico colocado sobre essa letra que indica o tipo de movimento de dedo.

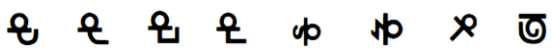
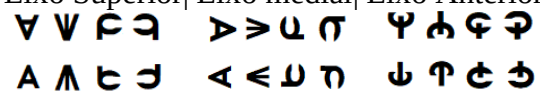
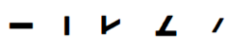
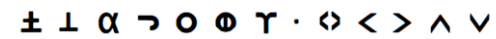
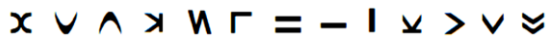
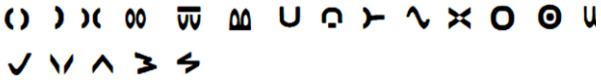
Quadro 15 – Caracteres da escrita Sel11

Letras de configuração de mão (CM)[12]	Letras de partes do corpo (PC)
	
Letras de movimento de mão (MovM)	Letras de representação dos cinco dedos e combinação de dedos (D e CD)
Movimentos retilíneos:  Movimentos em plano[13] Transversal:  Sagital: 	Os cinco dedos: Combinações dos dedos: 

[11] Um quadro com indicações do que representam cada caractere se encontra em anexo.

[12] Essas letras representam as configurações da mão direita. Quando invertidas horizontalmente, essas mesmas letras representam a mão esquerda. E há uma versão maiúscula delas para nomes próprios e início de frases.

[13] As quatro primeiras letras de movimento, em cada fileira, nos planos transversal e sagital e as sete primeiras no plano frontal (correspondentes respectivamente aos movimentos semicircular, curvo, angular duplo, angular, sinuoso, zigue-zague e diagonal) ocorrem em quatro posições a depender da direção. O que é algo bem intuitivo, percebido visualmente. Exemplo: O movimento semicircular transversal ocorre na direção: frente-direita  , frente-esquerda , direita-frente  e esquerda-frente . E os caracteres sinuoso ( e), zigue-zague ( e) e diagonal ( , , e) têm mais uma forma invertida, também indicando outra direção, respectivamente:  , , , ,  , , e .

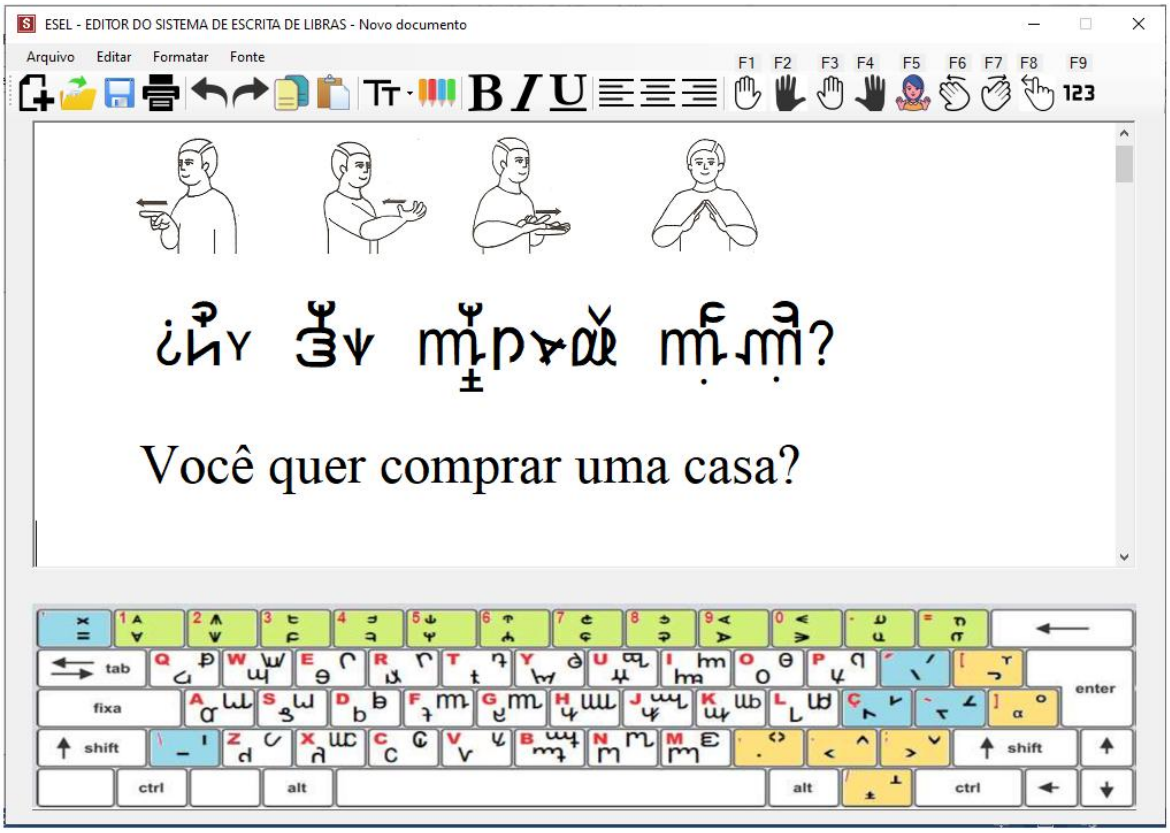
Frontal:  Movimentos fora de planos: 	
Diacríticos	
Eixo da mão e orientação de palma (colocados sobre as letras de CM) Eixos invertidos ➔	Eixo Superior| Eixo medial| Eixo Anterior 
Posição das duas mãos (colocados entre as letras de CM)	
Pontos de toque/proximidade (colocados sob as letras de CM)	
Movimentos de dedo(s) (colocados sobre as letras de D)	
Movimento/toque alternado ou conjunto (colocados entre as letras de MovM, D e PC)	
Movimento retilíneo brevíssimo e contido (colocados nas letras de Mov retilíneo)	
Expressões faciais (colocados sobre as letras de PC, ou MovM)	
Mudança de velocidade do movimento (mais lento ou mais acelerado) (colocados no final do sinal)	

Fonte: Elaborado a partir de Lessa-de-Oliveira (2023)

A Sel pode ser escrita com a utilização do Software Editor SEL, o E-SEL, que pode ser obtido gratuitamente.[14] Nesse editor, as letras e diacríticos são digitados utilizando-se o teclado físico. Pode-se transferir para o E-SEL imagens e também escrita com letras latinas. Ao instalar o aplicativo, as fontes da Sel são automaticamente instaladas. Assim, também é possível transferir o que é escrito em Sel no E-SEL para outros aplicativos (em computadores com o E-SEL instalado). Segue print screen da página do Software "Editor SEL":

[14] No momento, esse aplicativo está em fase de atualização para a versão mais recente da Sel que é a que foi publicada por Lessa-de-Oliveira (2023). Conforme informação dos responsáveis pelo aplicativo, dentro em breve a versão atualizada do E-SEL estará disponível para o público.

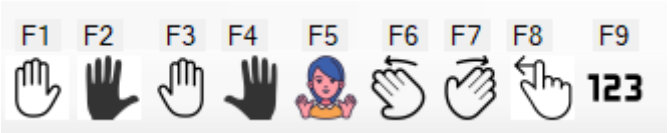
Figura 10 – Software “Editor SEL”



Fonte: Print Screen da página do Software “Editor SEL”

No canto direito da barra superior da face do E-SEL vemos ícones que indicam e dão acesso aos vários conjuntos de teclas que abrigam todos os caracteres da Sel, organizados da seguinte forma: F1, F2, F3 e F4 trazem as letras de configuração de mão e seus diacríticos; F5, as de partes do corpo e seus diacríticos; F6 e F7, as de movimento de mão e seus diacríticos; F8, as letras e diacríticos movimentos de dedos e sinais de pontuação; e F9 é um teclado numérico. O acesso aos conjuntos de teclas (ou teclados) pode ser feito clicando-se com o mouse nesses ícones indicados na barra superior, ou, de maneira muito mais fácil, apertando as teclas F de 1 a 9.

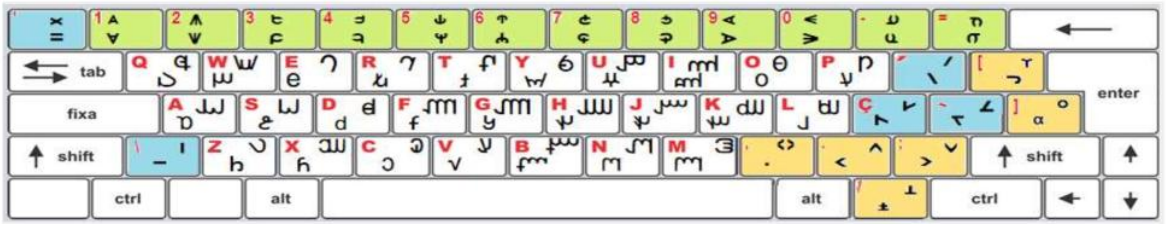
Figura 11 – Ícones de acesso aos conjuntos de teclas



Fonte: Recorte da face de E-SEL

A imagem do teclado que aparece na parte inferior da face do E-SEL é apenas um guia para auxiliar a identificação das teclas. Essa imagem muda sempre que se aperta as teclas F de 1 a 9 ou clica-se nos ícones da barra superior.

Figura 12 – Teclado F3 (configurações da mão direita minúsculas)



Fonte: Recorte da face de E-SEL

A acionarmos as teclas de F1 a F4 temos acesso às letras de configurações de mão nas suas quatro versões: F1 – minúsculas da mão esquerda; F2 – maiúsculas da mão esquerda; F3 – minúsculas da mão direita; e F4 – maiúsculas da mão direita. Nesses teclados, as teclas brancas trazem as letras de *configuração de mão*, as verdes trazem os diacríticos de *eixo-palma*, as amarelas os diacríticos de *toque/proximidade* e as azuis os diacríticos de *posição das duas mãos*. Ao acessar F1, F2, F3 ou F4 ocorre uma ligeira alteração apenas das teclas brancas. Em verdade, as letras não se alteram, muda-se apenas a versão da mesma letra, se minúscula, maiúscula, da mão esquerda ou direita, conforme a figura abaixo. Podemos observar que cada tecla apresenta duas letras. A letra que ocorre no canto superior é obtida utilizando-se a tecla Shift[15]

Figura 13 – Versões das CMs nas teclas brancas

Fonte: Recorte da face de E-SEL

Digitando F5, obtemos o teclado com as letras de partes do corpo nas teclas brancas e diacríticos de expressão facial (não-manual) nas teclas lilás. As teclas amarelas não se alteram e a tecla azul traz os diacríticos de alternância (.) e (..) conjuntamente.

[15] Tecla utilizada para se obter as letras maiúsculas no uso padrão dos teclados.

Figura 14 – Teclado F5 (partes do corpo)

Fonte: Recorte da face de E-SEL

Digitando F6 e F7, obtemos os teclados com movimentos de mão, nas teclas brancas. As teclas amarelas não se alteram e a tecla azul traz os diacríticos de alternância (.) e (..) conjuntamente.

Figura 15 – Teclado F6 (movimentos de mão)

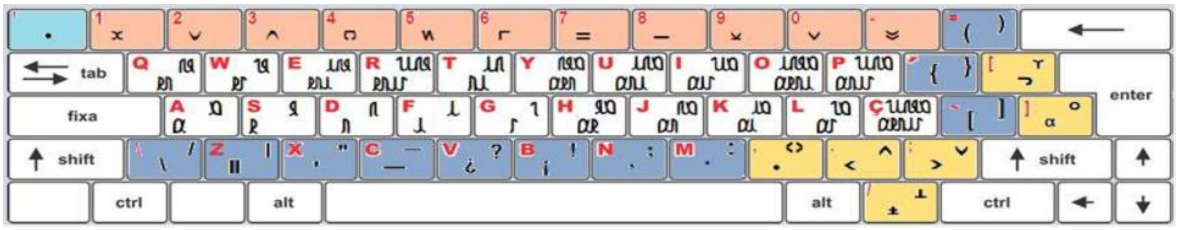
Fonte: Recorte da face de E-SEL

Figura 16 – Teclado F7 (movimentos de mão)

Fonte: Recorte da face de E-SEL

Digitando F8, obtemos o teclado com letras de dedos e combinações de dedos, nas teclas brancas, e diacríticos de movimentos de dedos, nas teclas rosas. As teclas amarelas não se alteram, a tecla azul celeste traz os diacríticos de alternância (.) e (..) conjuntamente e as azul turquesa trazem os símbolos de pontuação.

Figura 17 – Teclado F8 (movimentos de dedo(s))



Fonte: Recorte da face de E-SEL

Por fim, o E-SEL apresenta também um teclado numérico, ao qual se tem acesso digitando-se F9. Nesse teclado se encontram, além dos algarismos hindu-arábicos, os símbolos de dinheiro e números ordinais da Sel e outros símbolos matemáticos

Figura 18 – Teclado F9 (configurações da mão direita minúsculas)

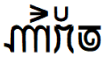
Fonte: Recorte da face de E-SEL

3.3.5 Comparação dos sistemas de escritas para línguas de sinais

Sintetizando, expomos, no quadro abaixo, uma síntese das principais características distintivas de cada sistema de escrita supracitado.

Quadro 16 – Síntese das principais características distintivas de cada sistema de escrita de língua sinalizada

SISTEMA DE ESCRITA	**SW**	**ELiS**	**VisoGrafia**	**Sel**
REPRESENTAÇÃO	Iconográfica	Arbitrária	Iconográfica	Arbitrária
ESTRUTURA DO SINAL	Mãos, Movimento, Expressão Facial e Corpo	Configuração de Dedos, Orientação da Palma, Ponto de Articulação e Movimento.	Configuração de Mão, Locação, Movimentos, Direção ou Orientação da Palma e as Expressões Não Manuais.	Mão, Locação e Movimento (com respectivos traços)
ORIENTAÇÃO DAS PALAVRAS	Linear de cima para baixo ou linear da esquerda para a direita	Linear da esquerda para a direita	Linear da esquerda para a direita	Linear da esquerda para a direita

ORIENTAÇÃO DOS CARACTERES NO SINAL	Distribuição em espaço bidimensional sem posições definidas	Linear da esquerda para a direita	Distribuição em espaço bidimensional sem posições definidas	Linear da esquerda para a direita para as letras com diacríticos colocados em posições específicas acima e abaixo ou entre as letras.
SOFTWARE PARA ESCRITA	SignPuddle Online	(Basta baixar a fonte True Type desenvolvida por Peixoto)	Não tem	Aplicativo E-SEL
EXEMPLO DE ESCRITA – SINAL GOSTAR				

Fonte: Elaboração própria

3.4 A escolha do sistema de escrita para anotação

Parte de nossa justificativa da escolha da Sel como modelo de escrita para compor o módulo de transcrição da iniciativa que propomos diz respeito a certos motivos pelos quais não escolhemos as outras três iniciativas de escrita – SW, VisoGrafia e EliS. Ou seja, a escrita Sel apresenta uma série de aspectos que dão a esse sistema duas características fundamentais – a precisão e a economia. Essas são duas propriedades que, segundo a autora, promovem as condições necessárias a um requisito primordial de um sistema de escrita, que possa funcionar na vida cotidiana, que é a automatização do processamento na leitura e na escrita. Com base nisso o sistema de escrita Sel "está submetido a uma regra geral tácita de economia, de acordo com a qual se procura evitar as informações excessivas, a fim de garantir a automatização do processamento" (Lessa-de-Oliveira, 2023, p. 102). Assim, explica a autora que, na primeira fase de elaboração da Sel,

> o critério que fundamentava o trabalho era o da busca de precisão na representação dos traços fonológicos do sinal, a partir daí o critério fundamental passou a ser o de buscar as condições para automatização do processamento na leitura e na escrita. Foi seguindo essas rotas que me foi possível chegar ao sistema apresentado neste livro, o qual, depois de mais de uma década de refinamentos, se mostra capaz de representar os sinais da Libras, ao mesmo tempo, com a precisão necessária à reprodução inequívoca da articulação do sinal escrito, no ato da leitura (mesmo que não se conheça o significado do item lexical escrito, o que é próprio de escritas alfabéticas) e capaz de proporcionar a automatização do processamento, que é o fator responsável, nas escritas em funcionamento no mundo, pela leveza de uma decodificação sem a necessidade de recorrência consciente às regras do sistema o tempo inteiro (Lessa-ee-Oliveira, 2023, p. 16-17).

Então, assumindo a importância das condições de automatização do processamento oferecidas pelos sistemas, fazemos uma análise desses quatro sistemas, procurando verificar o que favorece e desfavorece esse requisito. Começando pelo sistema SW, este se apresenta como um sistema fundamentalmente de característica icônica, o que o aproxima de composição de desenhos dos sinais, ao invés de um sistema baseado na fonética e fonologia da língua. Assim, como já mencionamos, na escrita dos sinais este é um sistema não linear, fincando os caracteres dispostos no plano bidimensional. E por que isso seria uma desvantagem? Porque uma disposição não linear dos caracteres dificulta a decodificação, dificultando o processamento na leitura, podendo comprometer sua automatização, uma vez que o leitor não encontra um caminho único de decodificação, que ocorreria se os caracteres estivessem organizados numa linha e a leitura se desse sempre na mesma direção a falta de um caminho de decodificação. Por não encontrar esse caminho automático, a decodificação não sai do plano consciente, o leitor não deixa de pensar nas regras do sistema, enquanto ler.

A falta de uma correlação fiel entre os caracteres do sistema os segmentos articulatórios constitutivos dos sinal é a segunda dificuldade séria que verificamos na SW. Isso foi mencionado atrás, quando observamos que a presença dos caracteres  , respectivamente nos sinais EU, QUER, e CASA, iconicamente nos remetem à ideia de uma cabeça/rosto com uma expressão facial, não correspondem a segmentos desses sinais, os quais não envolvem a cabeça, o rosto nem uma expressão facial em suas constituições articulatórias; e, também sem uma causa clara, nos sinais COMPRAR e UM, a dita ‘cabeça’ não aparece. Já o sinal EU apresenta um segmento não está representado por nenhum caractere nessa escrita – o *tórax*, tornando a escrita desse sinal semelhante, nesse aspecto, à escrita de uma sinal que realmente não apresenta uma parte do corpo como locação, como o sinal CASA. Tal inconstância também de compromete a automatização do processamento na escrita e na leitura.

O que parece haver é uma forte concentração da composição do sistema na iconicidade. Podemos perceber isso com mais clareza ao correlacionar a escrita dos sinais da frase “Eu quero comprar uma casa” com figuras da realização desses sinais, como na figura a seguir. Essa característica leva a um outro problema, o de ocupar muito espaço no papel. Como se trata de uma escrita bidimensional, ela ocupa aproximadamente quatro vezes o espaço de uma escrita linear.

Podemos apontar como uma terceira desvantagem desse sistema a dificuldade de reuso das tecnologias para escrita, uma vez que a aplicação não utiliza o teclado físico, é preciso

selecionar as imagens para formar a escrita dos sinais e frases, ou seja, é dependente da tecnologia do software "SignPuddle Online" o qual é específico para isso, o que acaba tornando-o uma tecnologia rudimentar. Analisamos que, provavelmente, a dificuldade em se elaborar um editor para o SW que possibilitasse a utilização do teclado físico para escrita esteja no fato de não se tratar de um sistema linear no nível na escrita interna dos sinais.

Passemos a um sistema semelhante ao SW, a VisoGrafia. Por ter tomado o SW, além do Elis, como base na sua elaboração, esse sistema apresenta, no geral, os mesmos problemas do SW. Assim, trata-se de um sistema fundamentalmente iconográfico, cuja escrita interna dos sinais se dá de forma não linear, em plano bidimensional, ocupando, no papel, quatro vezes o espaço de uma escrita linear. Dessa maneira, praticamente as mesmas dificuldades de automatização do processamento observadas no SW se encontram no VisoGrafia.

Além dessas dificuldades, parece não haver uma literatura mais ampla sobre a VisoGrafia disponível, que possa fornecer detalhes importantes sobre o funcionamento desse sistema, o que por si só torna o sistema inacessível. Além disso, o VisoGrafia não dispõe de uma proposta de software para escrita.

Quanto ao sistema da EliS, esse tem em comum com a escrita Sel a característica de se configurar como sistema arbitrário, baseado na fonética e fonologia da língua (com certa ressalva que mencionaremos adiante) e ser linear ao nível da escrita interna do sinal. A EliS também comunga com a Sel a característica de possuir estrutura de escrita facilmente lida pela máquina, pois, no caso da EliS, basta baixar a fonte True Type e, no caso da Sel, basta instalar o aplicativo E-SEL. Ou seja, em ambos os sistemas, temos fontes que se instalam no computador, criando a possibilidade de serem utilizadas em outros aplicativos; e, para ambos, o teclado físico é o recurso utilizado para a digitação.

Passando, todavia, para a questão do requisito de automatização do processamento, verificamos que o sistema da EliS apresenta, com base em nossos critérios de análise, certa inconsistência em relação à representação dos segmentos fonológicos naturais do sinal. O primeiro aspecto que, no nosso entender, leva a essa inconsistência diz respeito à utilização de *configuração de dedo* em lugar de *configuração de mão*. A configuração de mão como segmento fonológico do sinal é consenso entre os pesquisadores, seja como equiparável ao fonema, como propôs Stokoe (1960) entre outros, seja como traço, como propõe Lessa-de-Oliveira (2012; 2023). Quanto à configuração de dedo, parece não haver nenhuma pesquisa que tenha apontado isso. O que quer dizer que a configuração de mão é uma entidade psíquica presente no módulo fonológico da língua internalizada pelo falante; já a configuração de dedo não seria tal tipo de entidade. Esse aspecto pesa contra o fornecimento de condições à

automatização do processamento, pois se o caractere não representa um segmento fonológico natural, não há como uma operação cerebral remetê-lo a um dos segmentos presentes no módulo mental como entidades psíquicas.

Olhando para o próprio sistema da EliS, vemos isso com muita clareza. Primeiramente, verificamos que as CD, referentes a cada dedo, não podem ocorrer independes uma das outras. Elas têm que estar sempre se combinando para, em conjunto, remeterem ao que a gente compreende como uma configuração de mão. Afirma Barros (2016, p.205) que os "visografemas de Configuração de Dedos representam posições dos dedos e são combinados entre si para compor um Formato de Mão." Q quadro abaixo traz as configurações de dedo do sistema da EliS.

Figura 19 – Configurações de dedo do EliS

CONFIGURAÇÃO DE DEDOS	
Polegar	**Demais dedos**
. fechado	. fechado
⟋ na palma	⌝ muito curvo
⟍ curvo	⌝ curvo
⟍ "3D"	⟍ inclinado
_ horizontal	ǀ estendido
ǀ vertical	

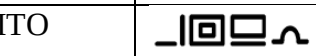
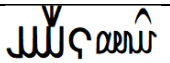
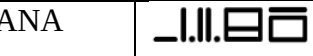
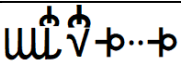
Fonte: Barros (2016, p. 205)

Algumas questões não parecem muito claras, nem muito intuitivas na combinação das configurações de dedos desse quadro. Por exemplo, não é muito intuitivo como diferenciar, nesse sistema,  (c) de  (o),  (efe) de  (t),  (argola espraiada) de  (cê-espraiado), entre outros. Outro problema com o processamento dessas configurações de dedos ocorre em sinais com mais de uma mão. Pelo que observamos da escrita do sinal BONITO, abaixo, a configuração de mão desse sinal que é  (mão espraiada) foi escrita como _ǀ ,com apenas um traço reto horizontal, para polegar, e um traço reto vertical, para os demais dedos, por esses estarem configurados da mesma maneira. Já na escrita do sinal SEMANA, as configurações de mão aparecem como _ǀ.ǁ.. As duas configurações de mão aí representadas são _ǀ  (mão espraiada) e .ǁ.  (vê). Entretanto, nada parece impedir que os três primeiros caracteres sejam processados como um bloco resultando na configuração _ǀ. 

(ele). Isso constitui um ruído para o processamento, ainda que o que sobrou ( ) não possa ser confundido com nenhuma outra combinação que resulte em outra configuração de mão, desfazendo a ambiguidade quanto aos três primeiros caracteres. É bem pouco provável que a automatização do processamento se dê nessas condições.

Na escrita Sel, o processamento das configurações de mão ocorre automaticamente porque cada configuração é representada por uma letra, ficando as da mão esquerda invertidas em relação às da mão direita, seguindo a anatomia, além de certa iconografia, que faz com que o formato dessas letras lembre o formato das configurações de mão –  ,  .

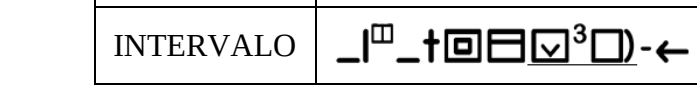
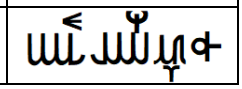
Quadro 17 – Os sinais BONITO e SEMANA em EliS e em Sel

	EliS	Sel
BONITO		
SEMANA		

Fonte: Elaboração própria

Outra dificuldade que verificamos na ELiS diz respeito à distribuição de caracteres sem observação de hierarquia entre os segmentos, distanciando os que deveria estar juntos porque precisam ser processados conjuntamente. É o que ocorre, por exemplo com o segmento *orientação de palma* em sinais bimanuais assimétricos. Assim, na escrita do sinal INTERVALO, em ELiS, entre os caracteres de configurações de dedo da mão não dominante ( ) e o caractere de orientação de palma referente à mão não dominante (-palma para trás) encontram-se os caracteres de configurações de dedo da mão dominante; e entre os caracteres de configurações de dedo da mão dominante ( ) e o caractere de orientação de palma referente à mão dominante (- palma para cima) encontram-se os caracteres de configurações de dedo da mão não dominante. Diferentemente, na Sel, os traços referentes a cada mão se encontram reunidos em uma letra e um diacrítico (- mão não dominante, - mão dominante). Claramente, as condições de processamento são bem melhores na Sel, nesse quesito.

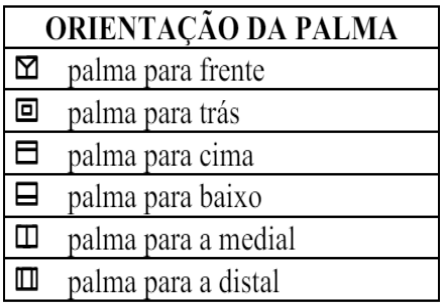
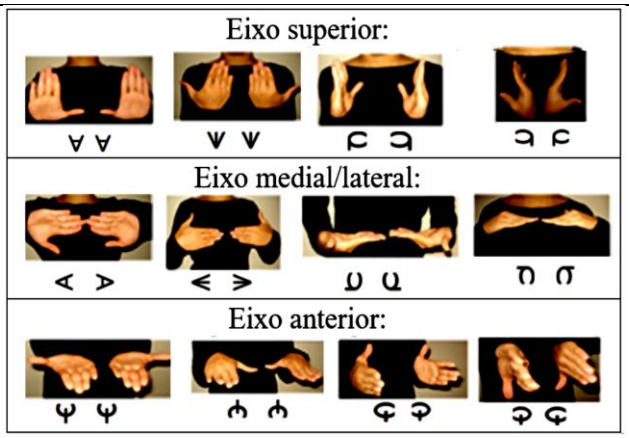
Quadro 18 – O sinaL INTERVALO em ELiS e em Sel

	ELiS	Sel
INTERVALO		

Fonte: Elaboração própria

Observamos que faltam na ELiS alguns traços representados na Sel. No quadro abaixo verificamos que enquanto na EliS o segundo grupo de caracteres representa apenas o traço orientação de palma, na Sel, o conjunto de diacríticos que são colocados obrigatoriamente sobre a letra de configuração de mão representam além do traço orientação de palma, também o traço eixo da mão. Cada eixo apresenta quatro orientações de palma, como se verifica nesse quadro, e cada eixo apresenta a sua versão invertida.[16]

Quadro 19 – Diferença de representação de traços entre EliS e Sel

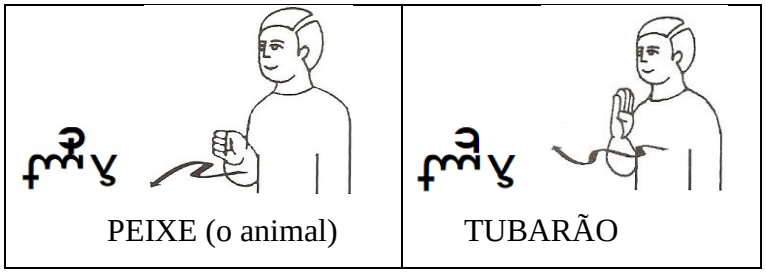
Orientação de palma na EliS	*Eixo da mão e orientação de palma na Sel*
	

Fonte: Elaborado a partir de Barros (2016) e Lessa-de-Oliveira (2023)

O traço de eixo da mão, que não vemos ser marcado na EliS, é distintivo como podemos verificar, por exemplo, observando o par mínimo PEIXE e TUBARÃO. Esse dois sinais da Libras se diferenciam apenas pelo eixo de posição da mão, que é anterior em PEIXE e superior em TUBARÃO, sendo a orientação da palma a mesma em ambos – para medial – como se pode observar na figura abaixo. A Sel dá conta dessa distinção com facilidade através dos diacríticos:  eixo anterior, palma para medial; e  eixo superior, palma para medial.

[16] Para mais detalhes, verificar a versão completa do quadro desses diacríticos da Sel em anexo.

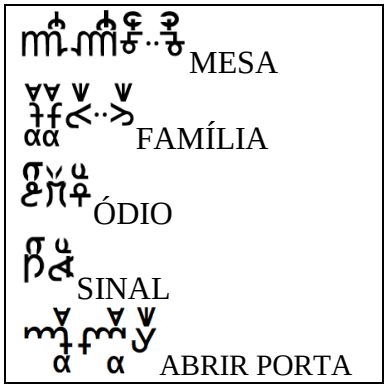
Figura 20 – Par mínimo pelo traço eixo da mão



Fonte: Elaboração própria com imagens de Capovilla *et al* (2011)

Em Sel esses diacríticos são utilizados também para marcar mudança de eixo e/ou orientação de palma durante o movimento de certos sinais, grafando-os sobre as letras configuração de mão e as letras de movimento, tal como nos sinais, abaixo:

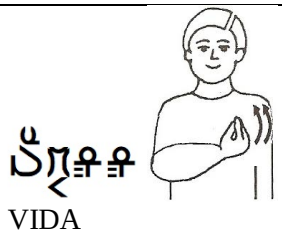
Figura 21 – Mudança de eixo e/ou orientação de palma durante o movimento em Sel

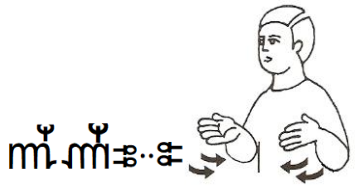
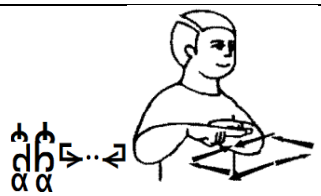
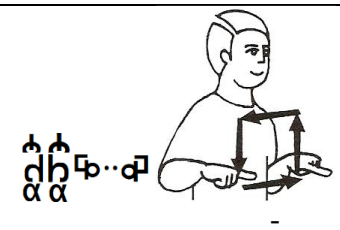
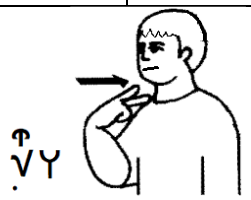
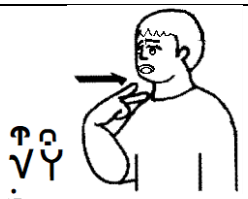
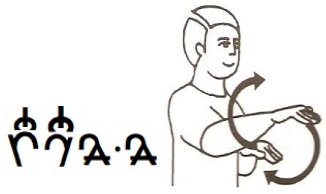


Fonte: Elaboração própria

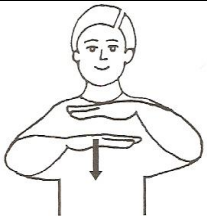
Além do traço de eixo da mão, percebemos, no sistema ELiS, a ausência de representação de outros traços que são igualmente necessários a uma escrita inequívoca, requisito fundamental à automatização do processamento na leitura e na escrita. Podemos verificar com clareza a importância da representação desses traços observando como em escrita Sel, esses são fundamentais na distinção de pares mínimos ou representação precisa do desenho icônico do sinal.

Quadro 20 – Exemplos de traços representados na Sel que estão ausentes na EliS

Descrição	**Sinais - pares mínimos ou análogos**	
Distintos apenas pelo traço Toque/proximidade	 VIDA	 PRESENÇA

Distintos apenas pelo aspecto do movimento retilíneo – neutro e brevíssimo –	 HOJE	 AGORA
Distintos apenas pelos traços de plano e direção de movimento	 QUADRA	 QUADRADO
Distintos apenas pelos traços plano e direção de movimento	 BRASIL	 TUBARÃO
Distintos apenas pelo traço de Expressão facial	 NERVOSO irritado	 NERVOSO ansioso
Distintos apenas pelo traço de Expressão facial	 NÃO PODE	 OCUPADO	 [17] gíria “SE FERROU”
Distintos apenas pelo traço de movimento das duas mãos alternado/conjunto	 NADAR	 CANGURU

[17] No caso das figuras com alteração do cabelo, isso significa que são referentes a sinais, ou variantes de sinais, não encontrados no dicionário Capovilla *et al.* (2011).

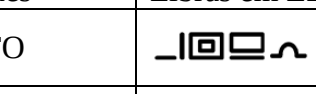
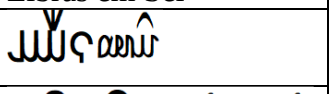
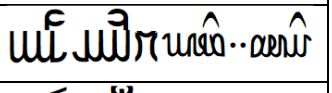
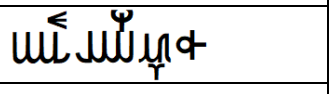
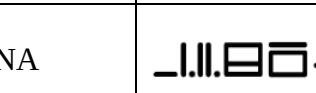
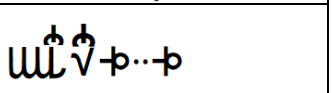
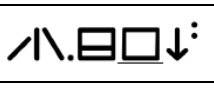
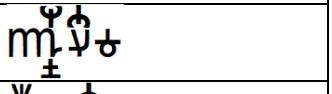
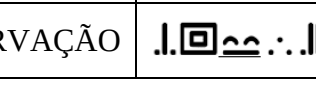
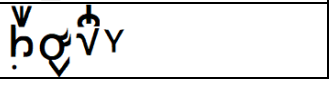
Distintos pelos traços de eixo e posição das duas mãos e direção do movimento.	 ABAIXO	 ACIMA

Fonte: Elaboração própria com imagens de Capovilla *et al* (2011)

Como podemos ver pelos exemplos aí apresentados, a representação de traços como eixo e posição das duas mãos, movimento das duas mãos alternado/conjunto, expressões faciais que compõem o sinal, plano de movimento, movimento retilíneo neutro, brevíssimo e contido, traços de toque amplamente definidos, pode ser necessária até à distinção de pares mínimos. Enquanto na ELiS a representação da maior parte desses traços está completamente ausente e outros estão timidamente representados, a escrita Sel mostra-se precisa na representação de todos eles, oferecendo também por isso melhores condições ao processamento.

Outra implicação para as condições de processamento na ELiS está relacionada a escrita de sinais realizados com as duas mãos. Segundo Barros (2016), há seis tipos de sinais:

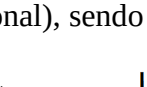
Quadro 21 – Comparação entre ELiS e Sel na escrita de sinais monomanuais e bimanuais

Tipo	Português	Libras em ELiS	Libras em Sel
Monomanual	BONITO		
Bimanural simétrico	LIMPO		
Bimanural assimétrico	INTERVALO		
Bimanural quase simétrico	SEMANA		
Com mão de apoio	VERDADE		
Composto	OBSERVAÇÃO		

Fonte: Elaboração própria.

Na ELiS há regras diferentes para escrita desses tipos de sinais, considerados por Barros (2016) como pertencentes a seis diferentes grupos. Os monomanuais seguem a regra geral de ordenação dos caracteres em: CD-OP-PA-M (Configuração de dedo, orientação de palma, ponto de articulação e movimento). Quanto aos bimanuais, no simétrico representam-

se as duas mão configuradas da mesma maneira e com a mesma OP, com duas barras inclinadas e uma única letra de OP; no assimétrico, ocorre um grupo de CDs referente a cada mão e, em seguida, uma letra de OP para cada mão; no quase simétrico, ocorre também um grupo de CDs referente a cada mão, mas apenas uma letra de OP para as duas mãos, que apresentam a mesma OP; já no grupo *com mão de apoio*, ocorrem apenas um grupo de CDs e uma letra de OP referentes à mão dominante e nenhuma CD e OP referentes à mão de apoio, que se configura e se posiciona diferentemente da mão dominante; por fim, no tipo *composto* ocorrem o grupo de caracteres de CDs seguido das letras de OP , PA e M, referente ao que para barros é o primeiro membro do composto, e em seguida ocorrem grupo de CDs, depois OP, PA e M do segundo membro do composto.

Essa grande variação de estrutura dificulta bastante as condições de processamento, preconizando enormemente a possibilidade de automatização do processamento. Na escrita Sel isto é resolvido de maneira muito simples e intuitiva. Nos sinais realizados com apenas uma mão, ocorre apenas a letra da configuração da mão principal com seus diacríticos de eixo-palma (como  em BONITO). Já nos sinais realizados com as duas mão, independentemente de haver mão dominante e mão passiva, ou as duas realizarem movimentos iguais ou diferentes, ou estarem em posições de eixo-palma iguais ou diferente, ocorrem um letra de |CM| para cada mão, com seu respectivo diacríticos eixo-palma acima (obrigatório) e toque/proximidade abaixo (ocasional), sendo a letra da mão esquerda (de base) sempre invertida em relação à da mão direita, como  em LIMPO, em INTERVALO, em SEMANA e em VERDADE. Dessa forma não há a menor necessidade de o falante se preocupar com o tipo de bimanual ao ler e escrever. E quanto ao último tipo, na Sel, este é compreendido como um sinal de duas unidades MLMov, isto é, de duas sílabas (1ª e 2ª), podendo ter sua constituição interno monomanual ou bimanual. Mesmo sendo sinais de fato compostos, isto é, com junção de mais de uma raiz semântica, em Sel são escritos justapostos, sem a utilização de hífen.

Frente a esses apontamentos, concluímos que o sistema SEL, até o presente momento, é a melhor escolha como sistema de escrita para compor o módulo de transcrição em nossa pesquisa, para ser a anotação fonológica do sinal que dá acesso ao sentido desse. A escrita Sel tem uma sistematização fonológica mais fiel à estrutura da Libras do que as demais, uma vez

que esta se baseia num modelo de estrutura fonológica hierárquica bastante interessante como vimos, o que a torna um sistema de escrita mais eficiente.

Portanto, a partir da percepção clara de que a escrita Sel[18] se mostra bem fundamentada e eficiente, atendendo melhor que as demais ao requisito de oferecer melhores condições de processamento por não apresentar os problemas que identificamos nas demais, sendo capaz de representar com precisão a estrutura fonológica do sinal, concluímos que ela é a melhor opção de escolha para nós, uma vez que, dessa maneira, ela atende satisfatoriamente ao que almejamos com o módulo de transcrição em nosso tipo de proposta de construção de corpora para línguas de sinais: anotação fonológica da Libras para ter acesso à realização do sinal.

[18] Todos os caracteres da escrita Sel em sua forma digital e manuscrita e as regras dessa escrita encontram-se expostos no anexo A e B, respectivamente. Encontram-se também, no anexo C, os caracteres da EliS mais completos.

4 CORPORA ELETRÔNICOS ANOTADOS E O LAPELINC FRAMEWORK

Uma vez que temos como objetivo central da pesquisa propor um workflow para construção de corpora de língua de sinais que reuse tecnologias e ferramentas já desenvolvidas para a construção de corpora de línguas orais, para além de verificar os limites e as possibilidades das iniciativas de construções de corpora de línguas de sinais existentes, faz-se necessário tratar de iniciativas de construção de corpora para línguas orais. Nesse sentido, na presente seção, apresentamos, de maneira sucinta, a relação entre a Linguística Computacional, a Linguística de Corpus, o Processamento de Linguagem Natural e de que maneira isso contribui para a pesquisa Linguística com base em corpora eletrônicos. Além de apresentar algumas metodologias já bastante fundamentadas e testadas para a compilação de corpora eletrônicos de línguas orais, enfatizando a anotação multicamada e a linguagem XML. Também apresentaremos de maneira suscinta os fluxos de trabalho ou sistemas de anotação do *Corpus Tycho Brahe* (CTB), do *Corpus* Kadiwéu, do *Corpus* Dialetal para o Estudo da Sintaxe (CORDIAL SIN) e do *Corpus* de Documentos Oitocentistas de Vitória da Conquista (*Corpus* DOVIC). Para finalizar, apresentaremos o *workflow* para a construção de corpora do método Lapelinc através do *Lapelinc framework* (Costa; Santos; Namiuti, 2021).

4.1 A Linguística Computacional, a Linguística de Corpus e o Processamento de Linguagem Natural

Costa (2015) salienta que Linguística Computacional é a área da Linguística que tem o objetivo de investigar o tratamento computacional da linguagem como também das línguas naturais para vários fins práticos, sendo um campo que se ocupa com o processamento por meio de computadores. Nesse sentido, a Linguística Computacional dialoga, em muitos sentidos, com (1) a Linguística de Corpus que é a área da Linguística responsável pela coleta e exploração de corpora, com o objetivo de realizar investigações sobre os fenômenos linguísticos e com (2) o Processamento de Linguagem Natural (PLN) que é "voltado para a construção de softwares, aplicativos e sistemas computacionais específicos, capazes de interpretar e/ou gerar informações em língua natural" (Costa, 2015, p.49)

Para explicar melhor, nas palavras de Othero (2006, p. 342-343):

> A Linguística de Corpus preocupa-se basicamente com o trabalho a partir de corpora eletrônicos que contenham amostras de linguagem natural. Essas amostras podem ser de diferentes fontes. Por isso, podemos encontrar os

> mais variados bancos de corpora eletrônicos: há corpora de linguagem falada, corpora de linguagem escrita literária, corpora com textos de jornal, corpora compostos exclusivamente por falas de crianças em estágio de desenvolvimento linguístico, etc. Os trabalhos envolvendo corpora linguísticos nem sempre têm como objetivo produzir algum software ou aplicativo. Normalmente, eles estão voltados para o estudo de determinados fenômenos linguísticos e sua ocorrência em grandes amostras de uma determinada língua (ou de uma variedade, dialeto ou modalidade dela). [...] A área de Processamento de Linguagem Natural, por outro lado, preocupa-se diretamente com o estudo da linguagem voltado para a construção de softwares, aplicativos e sistemas computacionais específicos, como tradutores automáticos, chatterbots, parsers, reconhecedores automáticos de voz, geradores automáticos de resumos, etc. Cabe à área de PLN justamente a construção de programas capazes de interpretar e/ou gerar informações em linguagem natural (Othero, 2006, p. 342-343).

Além disso, Othero (2006) ressalta a importância de se ter uma interação constante entre pesquisadores das áreas da Linguística e da Informática em trabalhos de Linguística Computacional. Isso porque, como afirma o autor, faltam ao linguista informações práticas e teóricas sobre linguagens de programação e desenvolvimento de softwares, ao passo que faltam ao engenheiro da computação saberes sobre teorias linguísticas para o trabalho com linguagens naturais. Sendo assim, a Linguística Computacional prospera com a cooperação e o diálogo entre essas duas áreas.

A possibilidade de se realizar um maior número de pesquisas em corpora eletrônicos foi permitida justamente pelo avanço na utilização da informática, sendo assim, “nos anos 90 muitos projetos de compilação de corpora surgiram no mundo todo e, atualmente, diversos corpora eletrônicos estão disponíveis para análise em várias línguas” (Costa, 2015, p.35). Esses corpora são distribuídos entre línguas como Português, Inglês, francês, espanhol, alemão, tcheco, chinês, entre outras. Para fins de conhecimento, na língua portuguesa, por exemplo, há vários corpora eletrônicos de destaque. Costa (2015) elenca oito, são: Corpus NILC/São Carlos, Corpora do Projeto Lácio-Web, Corpus Brasileiro, CETEMPúblico, CETENFolha, CRPC, PHPB-RJ e o Corpus Histórico do Português Tycho Brahe.

4.2 Compilação de corpora eletrônicos de línguas orais: etapa de anotação

Segundo Silveira (2008, p. 29), "Compilar – ou criar– um corpus é projetar e codificar uma coleção de documentos coletados dentro de determinados padrões ou exigências, para a realização de estudos linguísticos ou computacionais de aprendizagem de máquina". É nesse sentido que Costa (2015) afirma que a atividade de compilar um corpus eletrônico envolve

várias etapas como a coleta, a preparação, a segmentação (ou tokenização) e a anotação dos textos.

A autora explica que o processo de coleta visa reunir documentos textuais no formato eletrônico seguindo diretrizes elencadas para a composição do corpus. A demanda desse processo dependerá dos documentos de interesse, uma vez que se esses forem documentos físicos haverá a necessidade de digitação ou digitalização, ou ainda de transcrição de áudios ou manuscritos. Já na preparação ocorre a transformação de formatos de textos de maneira que estejam em um formato adequado para serem lidos pelas ferramentas, na maioria das vezes é o de texto puro (formato TXT). Ainda que já estejam em formato eletrônico, muitos documentos podem não estar no formato aceitável pelas ferramentas que serão utilizadas para serem lidos e processados e, por isso, precisa haver essa conversão. No processo de segmentação (ou tokenização) é onde é feito a divisão dos textos em menores unidades e a determinação de limites de palavras, sentenças e parágrafos. Existem ferramentas computacionais que executam essa etapa de maneira automática. Por fim, na anotação dos textos realiza-se a disposição de marcações, etiquetas ou tags, por meio de inserção automática, semiautomática ou manual de anotações. Essa etapa tem a finalidade de auxiliar o pesquisador na análise, de forma que ele consiga extrair do documento o máximo de informações possíveis. A partir de agora, abordaremos com maior ênfase o processo de anotação.

Costa (2015) explica que a anotação nos textos que compõem o corpus funciona, devido a inserção de marcações por etiquetas, como um pré-requisito para os trabalhos de análise de corpus, uma vez que possibilitam a realização de buscas automáticas, permitindo, assim, que o pesquisador recupere no texto eletrônico as informações como palavras, frases ou sentenças por correspondência de padrão. Isso acontece porque essas buscas automáticas estão diretamente ligadas ao formato de codificação e ao tipo de informação que integram a etapa de anotação no corpus. Essa possibilidade de realizar buscas automáticas é interessante, uma vez que “associadas a ferramentas computacionais permitem identificar e analisar padrões de uso da língua dentro do corpus dado” (Costa, 2015, p.56).

As anotações presentes nos textos de corpus eletrônicos se classificam em vários tipos e níveis, os quais, de acordo com Aluísio e Almeida (2006), inserem informações morfológicas, sintáticas, semânticas, discursivos etc. Aqui, trataremos mais precisamente sobre a anotação morfológica e sintática, pois estão mais ligadas ao nosso objetivo.

É chamada de Part-Of-Speech tagging (POS tagging) a anotação morfológica e, também, a anotação que além de informações morfológicas, traz informações sobre a

marcação das classes gramaticais das palavras, sendo assim uma anotação morfossintática. De acordo com Costa (2015, p. 57), “a especificação da anotação POS em morfológica ou morfossintática está diretamente relacionada ao conjunto de tags utilizado (tag set) e à atribuição destas pelo parser”.

Costa (2015) explica que é chamado de treebank o corpus que recebe anotação no nível sintático – entretanto, não é limitado a isso, atende quando se refere de maneira geral aos corpora gramaticalmente analisados em forma de árvore. Esse treebank se caracteriza pela reunião de textos ou conjunto de sentenças nos quais se marca a estrutura sintática dos constituintes por meio, convencionalmente, do uso de colchetes ou parênteses etiquetados. Mengel e Lezius (2000) afirmam que existem variadas formas de representar e anotar a estrutura sintática de um corpus, como: Tipster, Penn Treebank, Susanne, NeGra e diversos formatos para corpora fragmentados.

TIPSTER é uma arquitetura que determina algumas anotações padrão e atributos relacionados. Ela carrega a anotação separada do texto, associando a informação original ao elemento span. Para entender melhor, abaixo está um exemplo de frase anotada sintaticamente em Tipster.

Figura 22 – Frase anotada sintaticamente em Tipster

Text

```
Cyndi savored the soup.
|0...|5...|10..|15..|20
```

Annotations

Id	Type	Span Start	Span End	Attributes
1	token	0	5	pos=NP
2	token	6	13	pos=VBD
3	token	14	17	pos=DT
4	token	18	22	pos=NN
5	token	22	23	
6	name	0	5	name_type=person
7	sentence	0	23	constituents= [1],[2],[3].[4],[5]
8	parse	0	5	symbol="NP",constituents= [1]
9	parse	14	22	symbol="NP",constituents=[3],[4]
10	parse	6	22	symbol="VP",constituents=[2],[9]
11	parse	0	22	symbol="S",constituents=[8],[10]

Fonte: Grishman (1996, p. 268).

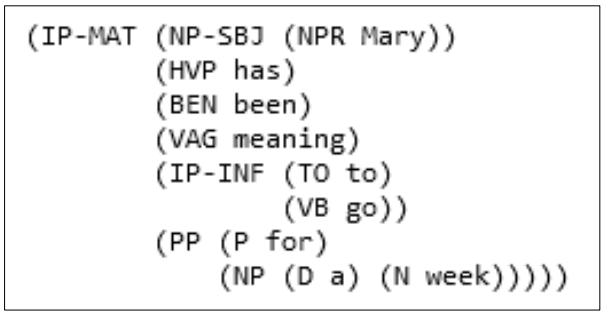
Na primeira linha da tabela está a frase que está sendo anotada; abaixo da frase há uma régua simplificada mostrando a posição (deslocamento de bytes) de cada caractere, a qual destrinchando, teríamos: |0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23|, sendo um número para cada caractere. Depois aparecem as anotações; uma

anotação por linha, determinadas pelas colunas de Id, Tipo, Spans e Atributos. Números inteiros são usados como IDs de anotação.

As anotações apontam para o fato que o programa conseguiu identificar na frase 5 tokens, os quais foram numerados de [1] a [5], definindo nos spans seu início e fim e em seguida seu atributo (ex: Cyndi=1:0-5:NP). O programa também organizou a hierarquia desses 5 tokens em [8], [9], [10] e [11], sendo que é em [11] onde encontramos a sentença maior composta por [8] e [10]: [[8:NP:Cyndi] e [10:VP: [2:VBD:savored] [9:[3: DT:the] [4:NN:soup]].

O Penn Treebank, por sua vez, utiliza uma representação limitada em árvore na forma de parênteses etiquetados. Segundo Santorini (2010), todos os parênteses abertos são acompanhados por um rótulo, seja esse de frase como NP – que assume posto de projeção máxima da oração – ou de palavra como N – que assume posto de núcleo da oração. Posto isso, segue, abaixo, um exemplo de frase anotada sintaticamente em Penn Treebank.

Figura 23 – Frase anotada sintaticamente em Penn Treebank



Fonte: Santorini (2010).

Santorini (2010) esclarece que as estruturas em Penn Treebank geralmente não incluem nem um VP nem projeções intermediárias como I'. Desse modo, o IP domina imediatamente todos os verbos e constituintes frasais.

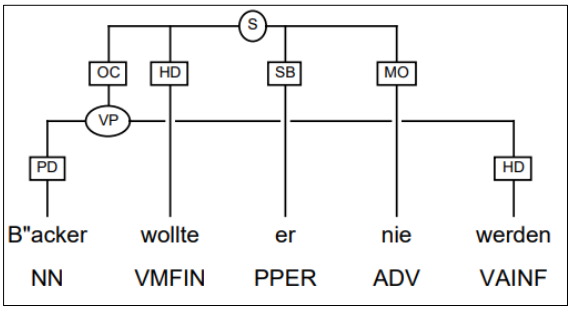
Já o Susanne, de acordo com Sampson (1995), é um esquema de anotação que produz uma estrutura constituinte etiquetada para qualquer sequência do Inglês, capaz de identificar de forma profunda todas as suas propriedades estruturais lógicas e superficiais, uma chamada "taxonomia lineana" para a língua inglesa.

Por fim, o NeGra, consoante Skut *et al*. (1998), é um esquema de anotação e ferramentas de anotação para texto alemão irrestrito. O formato de representação desse esquema é baseado principalmente na estrutura argumentativa, porém permite a extração de outras representações também. O autor afirma que para uma anotação rica, transparente e

consistente, os requisitos de descritividade, orientação aos dados e neutralidade teórica são frequentemente prezados.

Nesse modelo de anotação é utilizada a estrutura tectogramática – são estruturas de predicado-argumento que refletem a estrutura lexical do argumento e fornecem um guia para a montagem de significados – ao invés da estrutura fenogramática – reflete a ordem da superfície. Segue, abaixo, um exemplo dessa anotação:

Figura 24 – Frase anotada sintaticamente em NeGra



Fonte: Skut *et al*. (1998).

Além da estrutura tectogramática, a árvore também codifica informações de categoria sintática e anotações funcionais.

Como mencionado no início, além dessas, há diversas outras maneiras de se anotar sintaticamente a estrutura de corpora.

4.3 A linguagem XML

Consoante Costa (2015), a maioria dos padrões de anotação baseiam-se na linguagem XML (eXtensible Markup Language) para codificar internamente os documentos, a partir do uso de marcadores ou tags. Esse marcadores e tags não são fixos e nem limitados, os usuários têm a liberdade de estendê-los, criando seu próprio conjunto conforme sua necessidade específica. A XML é uma linguagem que oferece um padrão web universal e que funciona para determinar como um conteúdo vai ser apresentado na tela ou como os dados serão organizados dentro do documento codificado.

Apesar de se basearem em textos, a linguagem XML é capaz de descrever imagens, gráficos vetoriais, animações ou qualquer outro tipo de dado. Esses documentos são legíveis por pessoas e manipuláveis por computadores. Segue, abaixo, um exemplo de documento XML:

Figura 25 – Documento XML

```
<livro ISBN=”978-85-7244-800-0”>
   <autor> Carlos Mioto </autor>
   <autor> Ruth Lopes </autor>
   <autor> Maria Cristina Figueiredo Silva </autor>
   <titulo> Novo Manual de Sintaxe</titulo>
</livro>
```

Fonte: Costa (2015, p. 65).

A autora explica que este documento XML representa os dados de um livro e insere “as informações de autor, título e ISBN. O nó raiz é <livro> e este possui como filhos três nós <autor> e um nó <título>. A informação de ISBN foi representada como atributo do nó <livro> e seu valor no exemplo é “978-85-7244-800-0””. (Costa, 2015, p.65).

4.4 A anotação em exemplos de corpora

Como vimos, na construção de corpora eletrônicos temos a anotação gramatical que integra as etapas para compilação de corpus – a qual envolve informações morfológicas, sintáticas e semânticas – e temos a anotação em linguagem XML – a qual está relacionada a uma codificação interna dos documentos. A fim de compreender melhor o que foi exposto anteriormente sobre esses tipos de anotação, exemplificaremos o que foi explicitado até aqui com quatro corpora: *Corpus Tycho Brahe* (CTB), *Corpus* Kadiwéu, *Corpus* Dialetal para o Estudo da Sintaxe (CORDIAL SIN) e *Corpus* de Documentos Oitocentistas de Vitória da Conquista (*Corpus* DOVIC).

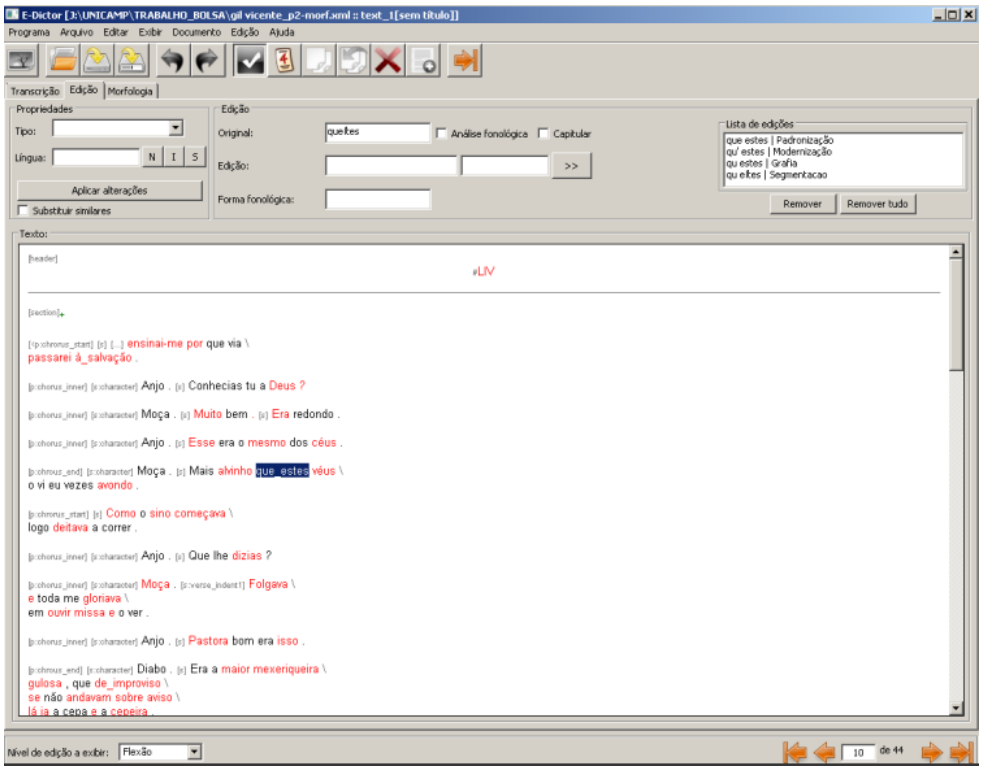
4.4.1 Corpus Tycho Brahe (CTB)

Segundo Galves (2019), o *Corpus* Sintaticamente Anotado do Português Histórico *Tycho Brahe* se caracteriza como um corpus eletrônico que tem a finalidade de armazenar dados que indiquem a história do Português, em relação à sua sintaxe. De acordo com Costa (2015), os textos que compõem esse corpus passam primeiramente pela etapa de transcrição, onde o arquivo é salvo no formato de texto simples (TXT) e em seguida passa pelas fases de edição e anotações morfossintática e sintática. Para transcrever, editar e anotar moforssintaticamente o *Corpus Tycho Brahe* foi elaborada a ferramenta “eDictor”, que é um editor de marcação extensível XML. Com ela é possível padronizar os textos a fim de rodar neles informações computacionais de anotação e busca e, além disso, continuar mantendo

acesso ao texto original – o que por sinal é imprescindível ao trabalhar com textos antigos como é o caso.

Galves (2019) esclarece que nessas etiquetagens uma correção manual é necessária, uma vez que elas são realizadas por um etiquetador automático probabilístico, treinado com dados do Português e sua taxa de acerto é de cerca de 95%.

Figura 26 – Tela de edição do eDictor



Fonte: Galves (2019, p.183).

A aba exibida na figura 14 apresenta a parte de edição do trecho do autor quinhentista Gil Vicente (como podemos ver a opção “Edição” selecionada no canto superior esquerdo da imagem). Galves (2019) explica que nela as palavras em vermelho que aparecem são as que foram modificadas e os traços entre palavras são os sinais da etapa de segmentação. Em “que_estes”, por exemplo, houve modificação, de modo que:

> A partir da forma queſtes (com o ſ da tipografia quinhentista), aplicam-se sucessivamente as operações de “segmentação”, que restitui duas palavras independentes (qu eſtes), de “grafia”, que substitui o tipo antigo ſ pelo moderno (qu estes), de “modernização”, que cria a sequência qu’estes, e enfim de padronização, produzindo a forma padrão hoje que estes (Galves, 2019, p. 183).

Costa (2015) elenca quatro anotações existentes no *Corpus Tycho Brahe:* (1) anotação da estrutura dos textos; (2) anotação de edições, (3) anotação morfossintática e (4) anotação sintática. Iremos explanar, de maneira sucinta, a respeito de cada uma.

Na anotação da estrutura dos textos é utilizada a linguagem XML, a qual torna possível que todas as anotações sejam preparadas e armazenadas em camadas em um único arquivo gerado pelo E-Dictor. O elemento <format> é utilizado para informações sobre a formatação, que podem ser do tipo capitular <format> t=”cap”></format>, do tipo itálico <format> t=”i”></format> ou do tipo negrito <format> t=”b”></format>. Os textos são divididos em parágrafos <p> e esses parágrafos são divididos em sentenças <s>. A presença de quebra de linha, quebra de página, quebra de coluna, capítulo, prefácio, prólogo, carta, índice, peça (teatral), ato, descrição dos personagens, marcação de cena, nome dos personagens, título, tabela e texto na margem são marcadas com uso da tag e diferenciadas entre si pelo atributo "t", como em <sec t=“line”/> para marcar quebra de linha ou em <sec t=“title”></sec> para marcar título. Informações sobre número da página, número da linha, número do parágrafo, cabeçalho da página, reclame de pé de página e imagem são anotadas fazendo uso da tag <text>, como em <text_el t=“par_nr”></text_el> para marcar número do parágrafo.

Na anotação de edições são feitas modificações nos textos do Tycho Brahe, visto que os textos antigos possuem características gráficas e grafemáticas que interferem no processamento computacional. Entretanto, como mencionamos anteriormente, para estudos filológicos, as características do texto original e ter acesso a esse texto original é importante. Sendo assim, na técnica de anotação adotada o texto é codificado com etiquetas XML para as estruturas variantes, mas o original é preservado, pois assim os textos podem ser recuperados de várias formas: em sua forma original ou com as edições realizadas.

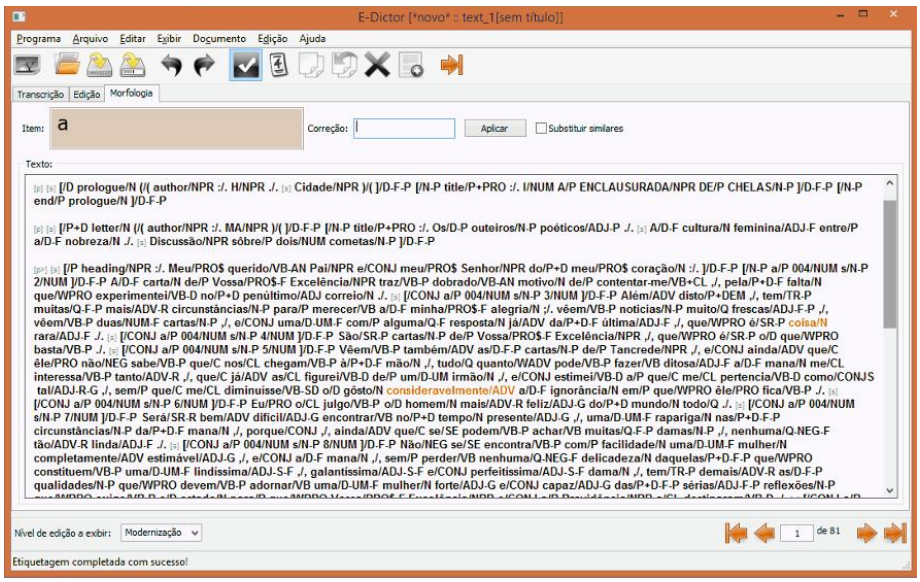
Desse modo, a anotação em linguagem XML é realizada identificando todas as interferências do editor sobre o texto original como uniformização grafemática, separação de vocábulos, junção de vocábulos, expansão de abreviatura, uniformização de pontuação, modernização de grafia e correções com elementos <e>, como em <e t=“jun”> ſe</e><o>queſe</o>, para marcar junção de vocábulos.

Na anotação morfossintática tem-se um sistema dividido em subníveis, formado por dois grupos de etiquetagem: o das etiquetas categoriais, que é usado na determinação do item lexical de acordo com a classe da palavra a que pertence; o das etiquetas flexionais, que acrescentam diacríticos para determinar ser de natureza verbal, designadores de informações modo-temporais ou não-verbal, indicadoras flexionais de gênero e número. Então, a etiqueta é

“composta por uma parte principal que indica a parte da classe POS (Part-of-Speech) ao qual o item lexical pertence, podendo ser acompanhada ou não de uma parte secundária, que especifica um subgrupo de determinada classe, ou vários traços flexionais carregados pelo item” (Costa, 2015, p. 81). O sinal diacrítico "-" é utilizado para conectar partes primárias e secundárias e "+" combina as classes POS ao se usar mais do que um, como em contrações.

Esse texto anotado em POS funciona como uma entrada para uma transformação na anotação XML. O que acontece é que o E-Dictor realiza a conversão para etiquetas na linguagem XML e as mantém no mesmo arquivo juntamente com as edições. Esse texto com etiquetas POS que foi editado no E-Dictor pode ser visualizado na ferramenta e um arquivo dele pode ser salvo no computador. Na figura 15 abaixo, retirada de Paixão de Souza; Kepler; Faria (2010), encontramos uma demonstração da anotação morfossintática com etiquetas POS no E-Dictor.

Figura 27 – Visualização de texto com etiquetas POS na ferramenta E-Dictor



Fonte: Paixão de Souza; Kepler; Faria (2010 citado por Costa, 2015, p.82).

Além desse formato de apresentação pela ferramenta trazemos outro exemplo de transcrição e etiquetagem morfossintática em formato somente texto, retirado de Galves (2019, p. 185):

a. Original:

“a prezentou huma Planta offerecida pello Ilustríssimo Senhor Jozé Corréa Machado arquiteto da Provincia, munto digno Socio Protetor da nossa Sociedade e por este mesno Senhor foi nos derigido o Conpetente Orcamento da Obra”

b. Versão etiquetada:

“apresentou/VB-D uma/D-UM-F planta/N oferecida/VB-AN-F pelo/P+D Ilustríssimo/ADJ-S Senhor/NPR José/NPR Corréa/NPR Machado/NPR arquiteto/N da/P+D-F província/N ,/, muito/Q digno/ADJ sócio/N protetor/ADJ da/P+D-F nossa/PRO$-F Sociedade/NPR e/CONJ por/P este/D mesmo/ADJ senhor/NPR foi/SR D nos/CL dirigido/VB-AN o/D competente/ADJ-G orçamento/N da/P+D-F obra/N”

Podemos verificar no exemplo acima a ilustração da aplicação dessas etiquetas à uma versão padronizada de uma sentença do texto “Atas dos Brasileiros”. Aqui, foram utilizadas também etapas de anotação de edição como a junção de palavras e também de padronização da ortografia.

Na anotação sintática é utilizada a anotação morfossintática para a aplicação de analisadores sintáticos ou “parsers”, que associam uma estrutura sintática a cada frase, isto é, a versão etiquetada dos textos é base para representação da estrutura sintática dos mesmos. São acrescentadas etiquetas de categorias sintagmáticas como IP-MAT, NP-ACC, NP-PRN, ADJP, PP. Abaixo, por exemplo, está a representação da estrutura sintática de b., retirada também de Galves (2019, p. 187-188):

```
( (IP-MAT (VB-D Apresentou)
        (NP-ACC (D-UM-F uma)
                (N planta)
                (ADJP (VB-AN-F oferecida)
                        (PP (P pel@)
                                (NP (D @o)
                        (ADJ-S Ilustríssimo)
                        (NPR Senhor)
                        (NP-PRN (NPR José) (NPR Corréa) (NPR Machado))
                        (NP-PRN (N arquiteto)
                                (PP (P da@)
                                        (NP (D-F @a) (N província))))
                        (, ,)
                        (NP-PRN (ADJP (Q muito) (ADJ digno))
                                (N sócio)
                                (NP-PRN (N protetor))
```

(PP (P da@)
(NP (D-F @a) (PRO$-F nossa) (NPR Sociedade)))))))))
(ID VA_002_SPL,19.113))

Todas essas anotações dão condições para que sejam feitas buscas por itens lexicais, classes de palavras e por estrutura sintática, de maneira que "grandes quantidades de dados podem ser exploradas de maneira automática e confiável" (Galves, 2019, p.182).

4.4.2 Corpus Kadiwéu

Galves, Sandalo, Sena e Veronesi (2017) propõe para o Kadiwéu, que é uma língua indígena polissintética do Brasil, uma extensão da anotação POS do *Corpus* Sintaticamente Anotado do Português Histórico *Tycho Brahe*. Eles têm essa iniciativa pois falta nos corpora das línguas nativas brasileiras a anotação, ou seja, a marcação das palavras para que seja possível de se realizar uma análise sintática – inclusive fato muito parecido como o que ocorre com a Libras. Além disso, a anotação morfológica é de suma importância para as línguas polissintéticas, tendo em vista que tornam possíveis as buscas de propriedades gramaticais codificadas pelos morfemas.

Os pesquisadores explicam que com o objetivo de dar conta da rica morfologia flexional do Português, o sistema de marcação POS dos Penn Parsed Corpora do Inglês histórico foi adaptado de maneira que as tags fossem articuladas, com uma base correspondente à categoria da palavra (VB, D, N, NPR, ADJ) e uma ou mais etiquetas secundárias que marcam propriedades morfológicas (-D, -UM-F, -F, -G). Entretanto, no Kadiwéu isso não funciona, pois a informação abarcada pela morfologia é excessivamente rica. No Kadiwéu, a correspondência entre forma e características pode ser marcada em uma única tag, com exceção de alguns casos de morfemas e supleção. Sendo assim, de acordo com Galves, Sandalo, Sena e Veronesi (2017), nessa língua, morfemas e palavras são tratados da mesma maneira e possuem o mesmo valor, pois ambos têm suas tags específicas. Segue um exemplo de marcação em Kadiwéu retirado de Galves, Sandalo, Sena e Veronesi (2017, p. 632):

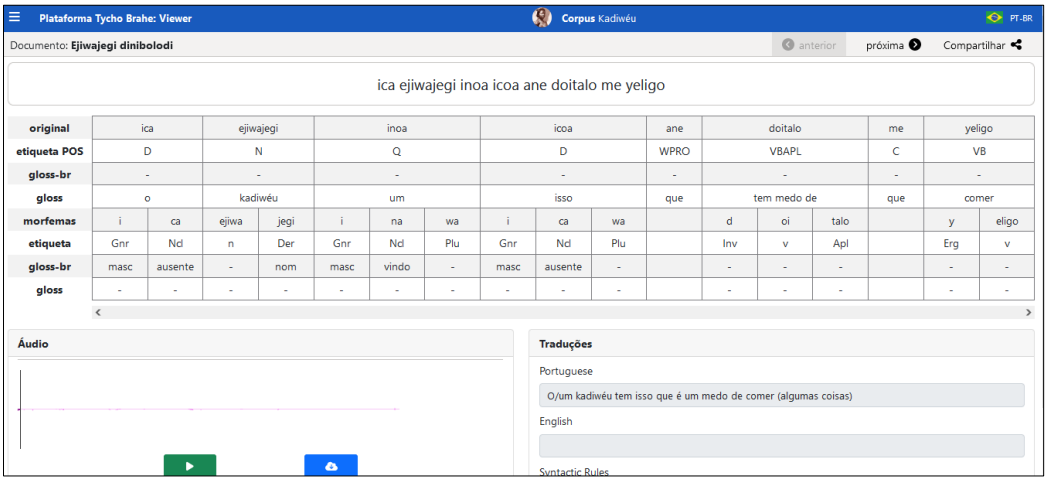
(1)	ijo	Gonel:egiwa	ja	wajipata.	
	i-jo	*Gonel:egi-wa*	*jaG*	*w-awajipa-ta-wa*	
	D	N		T	VB
	Gnr-Ncl	man-Cla		T	Erg-listen-Obl-Apl

'o/um homem ouviu isso' [19]

Nesse exemplo acima encontramos tags relacionadas a palavras como D, N e VB e também tags relacionadas a morfemas como Gnr, Ncl, man, Cla, Erg, listen, Obl, Apl. Como podemos observar, o processo de marcação em Kadiwéu se dá a partir de dois níveis: primeiro, o tagger é realizado no nível da frase e cada palavra recebe uma tag POS; depois, a divisão acontece dentro de cada palavra relevante e cada morfema recebe uma tag. Esse processo se difere do de idiomas como o Português por exemplo, pois nele, como vimos na seção 4.4.1, cada palavra recebe uma etiqueta que é composta por uma parte principal que carrega informações sobre a classe POS, acompanhada de uma parte segundária que pode especificar tanto um subgrupo de determinada classe como vários traços flexionais.

O *Corpus* Kadiwéu pode ser livremente acessado em https://www.tycho.iel.unicamp.br/viewer/C12 . Lá encontramos muitos exemplos de frases em Kadiwéu já anotadas. Na figura 16 a seguir trazemos a anotação de uma frase em Kadiwéu retirada do próprio site do corpus.

Figura 28 – Exemplo de frase anotada no Corpus Kadiwéu



Fonte: Print Screen do site do *Corpus* Kadiwéu

Como vemos, o *Corpus* Kadiwéu é organizado da seguinte maneira: a frase em cima em Kadiwéu; logo abaixo as anotações de etiqueta POS, gloos-br, gloss, primeiro de frase e depois de morfemas; embaixo o áudio com a reprodução da frase e ao lado a tradução da mesma.

[19] Tradução da autora. No original: ‘the/a man has listened to it.’

4.4.3 Corpus Dialetal para o Estudo da Sintaxe (CORDIAL SIN)

O CORDIAL SIN se caracteriza, em consonância com Magro e Vaamonde (2019, p. 252), como "um corpus geograficamente representativo dos dialetos do PE composto por excertos de discurso espontâneo selecionados a partir de gravações de inquéritos dialetais realizados em 42 localidades do território Português", ou seja, um corpus de variação linguística do Português europeu. Segundo os autores, o CORDIAL SIN se desenvolve satisfatoriamente no que diz respeito ao seu plano linguístico, no entanto, com relação ao plano tecnológico, devido à época em que foi construído, apresenta limitações que dificultam uma exploração suficiente dos seus dados.

Por essa razão, Magro e Vaamonde (2019) esclarecem que será implementado nesse corpus o projeto Atlas Sintático do Português Europeu – SynAPse. Com esse projeto, que auxilia no cruzamento de dados sobre o domínio da mineração de florestas sintáticas e representação cartográfica linguística, pretende-se construir ferramentas que levantem automaticamente resultados de estruturas sintáticas no CORDIAL SIN, com a conversão do corpus em linguagem XML.

Sendo assim, abaixo está esquematizado o processo de codificação do *Corpus* CORDIAL-SIN em linguagem XML, retirado de Magro e Vaamonde (2019, p.261):

1. Conversão dos 42 documentos com a transcrição conservadora em formato Microsoft Word (.doc) em 42 documentos em texto simples Unicode (.txt).

2. Criação de um arquivo XML para cada sequência de inquérito do CORDIAL-SIN. O número total de arquivos XML que constituem o corpus é de 2058. Cada arquivo XML está dividido em metadados (<teiHeader>) e texto (<text>).

3. Conversão dos metadados.

4. Tokenização do texto.

5. Conversão das marcas textuais de transcrição.

6. Importação da anotação morfossintática.

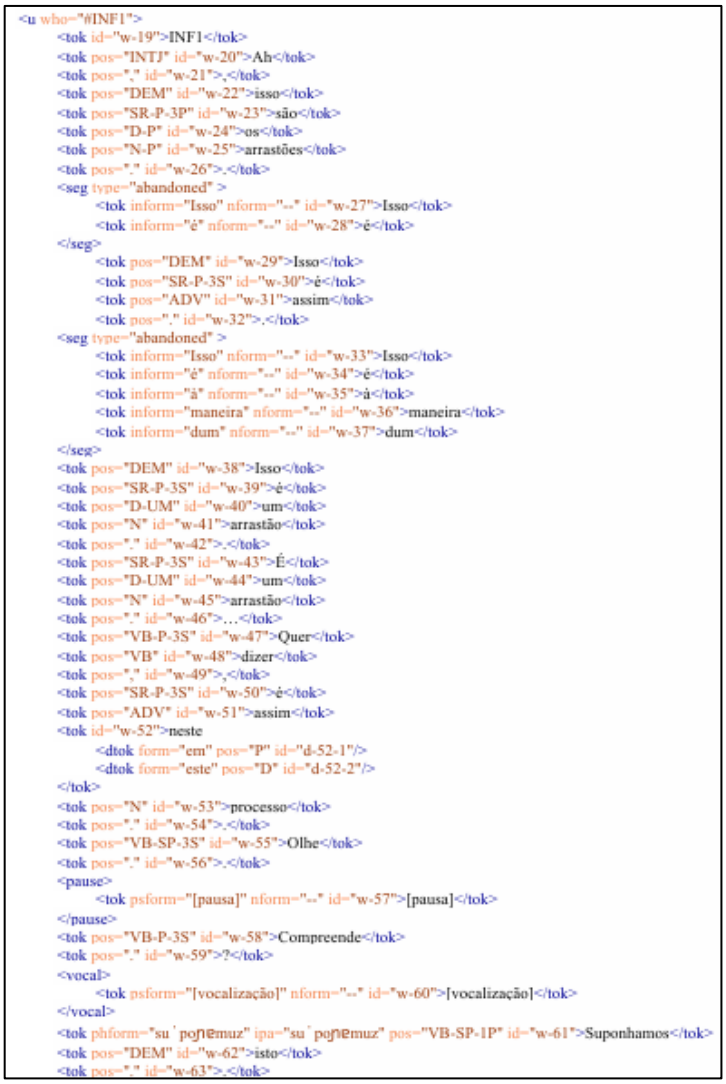
Sobre esse processo de conversão do corpus, Magro e Vaamonde (2019, p.261) esclarecem que:

> [...] foram adotadas as diretrizes de codificação propostas pelo consórcio TEI (Text Encoding Initiative) para a edição de textos em formato digital (TEI CONSORTIUM, 2019). Os standards TEI foram fundamentalmente aplicados à informação metatextual de cada ficheiro XML, isto é, à

> informação incluída no cabeçalho (<teiHeader>). Para a marcação de aspetos textuais, incluindo a própria tokenização do texto e a sua anotação linguística, adotou-se a estratégia de marcação utilizada pela ferramenta TEITOK (JANSSEN, 2014, 2016). O TEITOK é uma plataforma web especialmente desenhada para ver, criar e editar corpora que combinam marcação textual e anotação linguística, que é utilizada pelo projeto SynAPse para visualizar, editar e explorar os dados do CORDIAL-SIN uma vez codificados em XML (Magro; Vaamonde, 2019, p. 261).

Posto isso, trazemos em seguida, um fragmento de texto do CORDIAL SIN codificado no projeto SynAPse:

Figura 29 – Fragmento de texto em SynAPse



Fonte: Magro e Vaamonde (2019, p. 263).

Uma breve explicação sobre a figura 16 é que como Magro e Vaamonde (2019) elucidam cada token é marcado pelo elemento <tok> e cada elemento desse está associado a um identificador único por meio do atributo @id e a uma etiqueta morfossintática pelo

atributo @pos. Em situações em que uma palavra ortográfica se refere a duas ou mais palavras gramaticais, é utilizado <dtok> dentro de <tok>. Isso ocorre principalmente em casos de contrações e enclíticos.

4.4.4 Corpus de Documentos Oitocentistas de Vitória da Conquista (Corpus DOVIC)

O *Corpus* DOVIC, em consonância com Santos e Namiuti (2019), foi construído para ser um corpus digital anotado de documentos da região sudoeste da Bahia, no qual se procurou solucionar o problema concernente à fidedignidade entre o Documento Físico (DF) e sua versão de Documento Digital Texto (DDT). Para isso, nesse corpus aplica-se um método de construção de corpora digitais anotados cientificamente controlados que de um DF, gera-se um Documento Digital Imagem (DDI), o qual servirá de suporte no processo de construção de corpora anotados, e que depois se tornará um DDT com módulos de anotação especificando o processo realizado. Desse modo, na transposição material do papel para o suporte digital encontra-se um mecanismo que recupera, no digital, a complexidade do documento físico, o qual é chamado de Aparato de Metadados Estruturados (AME). O AME se dá a partir de cinco componentes: Catálogo Visual; Dossiê de Observações Pertinentes (DOP); Fotografia Cientificamente Controlada (FCC); Análise Topográfica (AT) e Análise Descritiva (AD).

Para entendermos melhor como funciona esse processo de criação e anotação do *Corpus* DOVIC segundo Santos e Namiuti (2019), segue, abaixo, ilustrações de cada componente do AME:

- Catálogo Visual

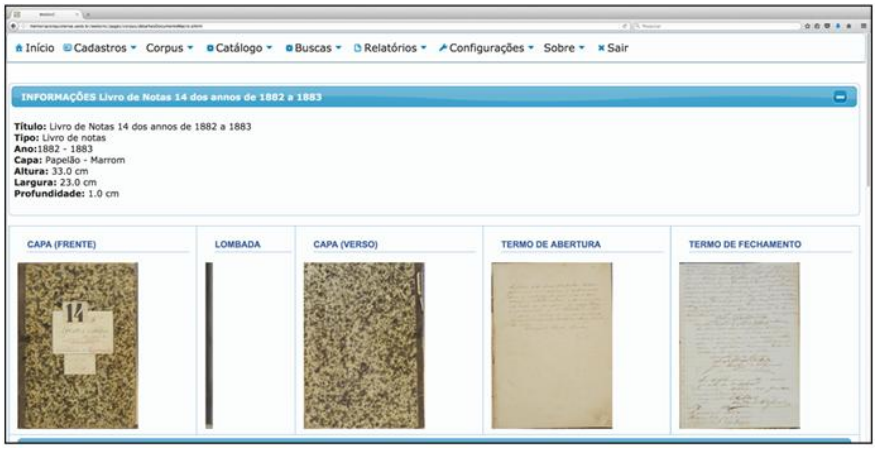
O Catálogo Visual é um produto gerado a partir do AME e funciona como base para a construção de filtros a partir de dados nele registrados.

Figura 30 – Recorte da tabela de dados da parte descritiva do catálogo visual referente aos livros de notas oitocentistas de Vitória da Conquista

Código LapeIinc	LIVRO No.	ANO	TIPO	TAMANHO (cm)			CAPA	DDI	Observação	Catalogador
				Altura	Largura	Profundidade				
			ESCRITURAS							
E1	1	1841-1848	Escrituras	31,5	22,3	3	Marrom	sim	Fita adesiva na lombada	Giovane, Adilson e Jorge
E2	2	1841-1855	Escrituras	32	22	3	Bege	sim	Fita adesiva na lombada	Giovane, Adilson e Jorge
E3	3	1849-1858	Escrituras	31	23	5	Capa couro	sim		Vanessa, Giovane, Jorge
E5	5	1852-1866	Escrituras	34	23	2	Capa Papelão	sim		Vanessa, Giovane, Jorge
E6	6	1869-1870	Escrituras	33	22	1	Azul	sim	Fita de tecido na lombada	Giovane, Adilson e Jorge
E7	7	1864-1871	Escrituras	34	23	3	Capa Papelão	sim		Vanessa, Giovane, Jorge
E8	8	1870-1874	Escrituras	33	23	1,5	Capa Papelão	sim		Vanessa, Giovane, Jorge
E11	11	1877-1880	Escrituras	32	23	0,5	Capa Papelão	sim		Vanessa, Giovane, Jorge
E13	13	1880-1881	Escrituras	33	23	1	Capa Papelão	sim		Vanessa, Giovane, Jorge
E14	14	1882-1883	Escrituras	33	23	1	Capa Papelão	sim		Vanessa, Giovane, Jorge
E15	15	1881-1882	Escrituras	33	23	1	Capa Papelão	sim		Vanessa, Giovane, Jorge
E16	16	1883-1884	Escrituras	31	22	1	Capa Papelão	sim		Vanessa, Giovane, Jorge
E21	21	1880-1890	Escrituras	47	33	4	Capa Papelão Preta	sim		Vanessa, Giovane, Jorge

Fonte: Santos e Namiuti (2019, p.1391)

Figura 31 – Tela de visualização do Catálogo Visual gerado pelo WebSinC exibindo elementos da parte descritiva combinados com as imagens-chave da parte imagética do Livro E 14



Fonte: Santos e Namiuti (2019, p.1392)

As imagens acima apresentam as duas partes do Catálogo Visual: a primeira a parte descritiva que serve para destacar as principais características do documento a fim de alimentar o banco de dados e a segunda a parte imagética que são imagens selecionadas para contribuir com a identificação do documento.

- Dossiê de Observações Pertinentes (DOP)

O DOP antecede a captura fotográfica. É onde ocorre a pré-análise das fontes originais registrando anotações de tipo fotográfica, filológica, e outras que possam ser relevantes para a captura imagética da fonte e/ou para a edição do documento em ambiente digital.

Figura 32 – Recorte da tabela de dados de um Dossiê de Observações Pertinentes

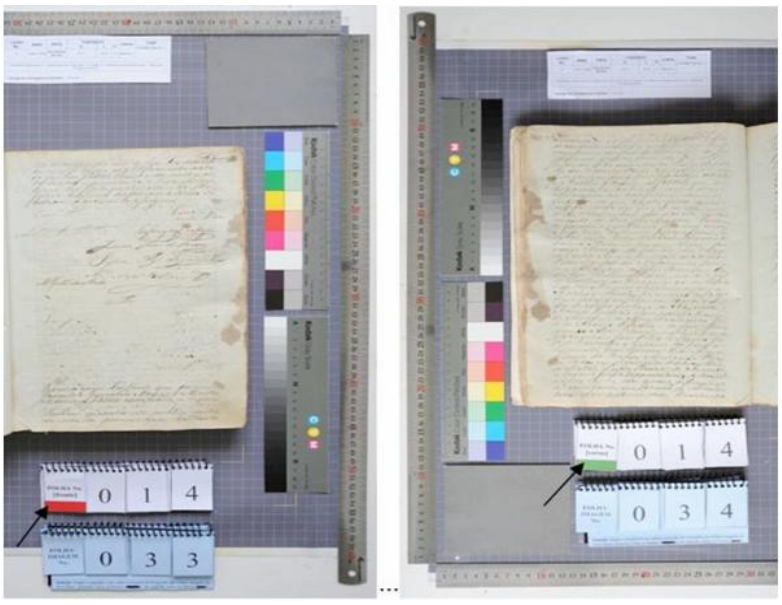
LIVRO No.	ANO	TIPO	TAMANHO			CAPA						
			A	L	P		CÓDIGO LAPELINC	Imagens	CÓDIGO DE OBSERVAÇÃO	FOLHA-IMAGEM	Tipo de Observações	OBSERVAÇÕES PARTICULARES (OP-XX)
1	1841-1848	Escrituras (Notas)	31,5	22,3	3	Marrom	C11-E01	412	34	Contracapa frente	Filológica	Folha anexada posteriormente. Transcrição do termo de abertura.
1	1841-1848	Escrituras (Notas)	31,5	22,3	3	Marrom	C11-E01	412	35	S/N	Filológica	Folha solta.
1	1841-1848	Escrituras (Notas)	31,5	22,3	3	Marrom	C11-E01	412	27	Termo de abertura	Filológica	
1	1841-1848	Escrituras (Notas)	31,5	22,3	3	Marrom	C11-E01	412	10	198	Filológica	Salta de 198 para 201
1	1841-1848	Escrituras (Notas)	31,5	22,3	3	Marrom	C11-E01	412	22	203	Filológica	Termo de encerramento com nome De: "Termo de Apresentação"
1	1841-1848	Escrituras (Notas)	31,5	22,3	3	Marrom	C11-E01	412	22	204	Filológica	Termo de encerramento com nome De: "Termo de Apresentação"

Fonte: Santos e Namiuti (2019, p.1393)

- Fotografia Cientificamente Controlada (FCC)

É onde ocorre de fato a captura fotográfica da imagem do original. São utilizados equipamentos adequados e inseridos dados necessários para garantir a relação da imagem com o objeto que a originou.

Figura 33 – Visualização da FCC do livro de notas E14 ordenado não editado com metainformações cientificamente controladas nas folhas-imagem dos DDIs, conforme o método LAPELINC



Fonte: Santos e Namiuti (2019, p.1393)

- Análise Topográfica (AT)

A AT marca a localização topográfica como folha-imagem inicial e final, e realiza uma identificação tipológica a partir de cabeçalhos dos documentos.

Figura 34 – Recorte da tabela de dados da Análise Topográfica

Documento Número	Cabeçalho	Tipo	Folha-imagem inicial	Folha-imagem final
01	Acta dos [ins...lhação?] do collegio	Ata	005	010
02	Carta de liberdade do Escravo Constatino conferida por seu Senhor Jorge de Oliveira Freitas como abaixo se declara.	Carta de liberdade	010	012

Fonte: Santos e Namiuti (2019, p.1394).

- Análise Descritiva (AD)

A AD se dá pela descrição linguístico-jurídica da tipologia dos documentos, a partir do auxílio de dicionários de língua históricos e contemporâneos, além de dicionários técnicos, dicionários específicos da escravidão e jurídicos.

O AME é apenas um dos aparatos desenvolvidos pelos autores para garantir a fidedignidade do documento digital em relação ao físico, tal aparato integra as ferramentas do método de construção de corpus do Laboratório de Pesquisa em Linguística de Corpus, o método LAPELINC (Santos; Namiuti, 2019).

O método LAPELINC possui um workflow que integra três grandes etapas para a construção de corpora: (i) transposição; (ii) transcrição; (iii) compilação.

A etapa de transposição envolve a captura em mídia eletrônica/digital dos dados/documentos a partir das fontes originais, no caso de documentos históricos do DOViC, livros manuscritos em papel que passam pelo processo de captura imagética.

A etapa de transcrição envolve a transcrição do texto (dado linguístico) em ambiente digital seguindo um sistema de anotação em camadas ligado aos objetos (imagens co-indexadas) da etapa anterior.

A etapa de compilação envolve outros tipos de anotação, como anotações de edição do texto e camadas de anotação da morfologia e da sintaxe do texto.

Consideramos que o workflow do método Lapelinc, elaborado por Santos e Namiuti para o corpus DOVIC, pode ser adaptado para a construção de corpora de língua de sinais e por isso trataremos mais dele na próxima subseção a partir de Costa, Santos, Namiuti (2021).

4.5 O Lapelinc Framework

De acordo com Costa, Santos e Namiuti (2021), é possível elencar um conjunto de etapas comuns no processo de construção de corpora, ainda que essas pesquisas tenham em si suas peculiaridades que requerem adaptações. Foi pensando nisso que os autores propuseram uma metodologia para se obter um padrão na criação de corpora linguísticos: o LAPELINC Framework.

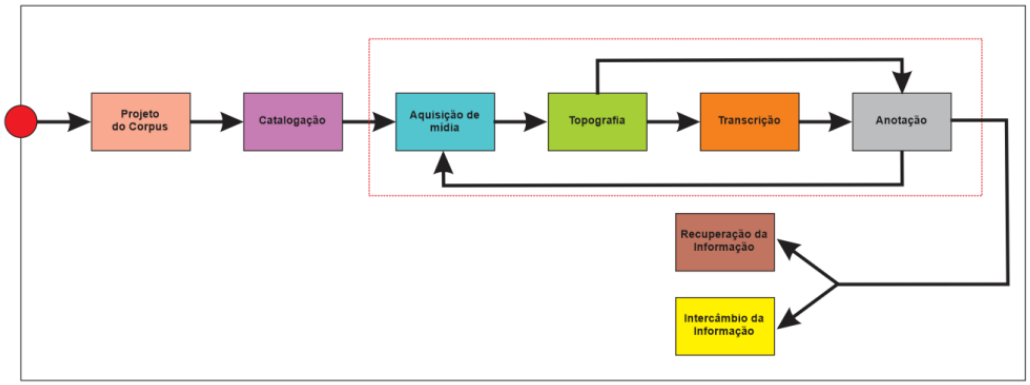
Para Costa, *et al*. (2022):

> [...] o Lapelinc Framework permite a criação de corpora de múltiplos tipos e finalidades, consistindo tanto em um conjunto de fluxos de trabalho que implementam uma proposta padrão para a construção de corpora quanto em um conjunto de ferramentas de software que permitem a implementação de corpora de acordo com a norma proposta (Costa, *et al*. , 2022, p. 403)[20]

Desse modo, se entende então que a partir do LAPELINC Framework se obtém um conjunto de etapas e subprodutos comuns da pesquisa linguística que tem o corpus como objeto. Segundo os autores, essa padronização de técnicas e procedimentos é interessante, uma vez que torna viável a possibilidade de resolver, de maneira também padronizada, problemas que tenham a mesma tipologia ou características semelhantes e, além disso, ainda reduz as margens de erro e chances de retrabalho.

Posto isso, abaixo está o esquema das etapas e o fluxo entre elas apresentadas pelo LAPELINC Framework.

Figura 35 – LAPELINC Framework



Fonte: Costa, Santos e Namiuti (2021).

[20] Tradução da autora. No original: the Lapelinc Framework allows the creation of corpora of multiple types and purposes, consisting of both a set of workflows that implement a standard proposal for the construction of corpora and a set of software tools that enable the implementation of corpora in accordance with the proposed standard.

Como podemos observar na figura 11 e como explicam Costa *et al*. (2022), cada estágio consiste em pontos fixos, chamados por eles de pontos congelados, que são as diretrizes indispensáveis; e pontos flexíveis, chamados por eles de pontos quentes, que são responsáveis por proporcionar a customização necessária para atender aquilo que é peculiar de cada pesquisa, de cada corpus que venha a ser construído.

A seguir, será discorrido, de maneira geral, com base em Costa, Santos e Namiuti (2021), acerca de cada etapa que compõe o LAPELINC Framework. Entretanto, antes disso, os autores esclarecem que:

> A etapa inicial é a de Projeto do Corpus. As demais etapas do Framework a serem cumpridas dependem necessariamente das demandas particulares de cada pesquisa. O framework admite ainda o trabalho de forma linear, isto é, cada etapa é executada uma após a outra e uma única vez, ou de forma interativa e incremental, permitindo que o conjunto de etapas seja executado o número de vezes que for necessário (Costa; Santos; Namiuti, 2021, p. 9).

1. Projeto do Corpus;

Nessa primeira etapa, ocorre o planejamento do corpus, a partir da definição de escopo e definição de modalidade ou tipo de corpus.

2. Catalogação;

É onde se dá a escolha pelo material com o qual se pretende utilizar no corpus, podendo realizar o levantamento de informações relacionadas a esse material.

3. Aquisição de mídia;

Nessa etapa, é feita a coleta e reunião do material a ser utilizado, a partir de cinco rotinas distintas para atender aos tipos existentes de corpus: transposição, aquisição de imagens digitais, aquisição de áudio, aquisição de vídeo e aquisição de texto em formato digital.

4. Topografia;

Aqui é realizada a anotação da estrutura topográfica, das características dos materiais, a partir de três rotinas: cadastro de documentos macro, cadastro de documentos micro e compilação de sumário.

5. Transcrição;

É onde são utilizados recursos e ferramentas específicos para a transcrição em texto das imagens, vídeos ou áudios.

6. Anotação;

Nessa etapa, é onde se dá o registro de metainformações de edição, morfossintaxe e semântica relacionados ao texto do corpus, a partir dos módulos de POS Tagging, Parsing Sintático, edição (e aqui que se encaixa a tradução, de acordo com os autores) e anotação semântica.

7. Recuperação da Informação;

Aqui é possível de se realizar operações de busca automática e visualização nos textos do corpus, a fim de se obter informações linguísticas.

8. Intercâmbio da Informação;

Nessa última etapa, é onde é possível de se fazer, utilizando tecnologias de serviço WEB, a importação/exportação de textos do corpus para diálogo com outras iniciativas de pesquisa, o que permite o compartilhamento de informações com a comunidade científica.

Essa metodologia – o LAPELINC Framework – é parte do trabalho de tese de Costa (2019) e foi elaborada baseada no fluxo de trabalho LAPELINC para construção de corpora históricos. Entretanto, ainda que ela tenha como base o fluxo de trabalho LAPELINC para construção de corpora especificamente históricos, é, como já foi mencionado, um framework customizável capaz de apresentar um padrão de trabalho para qualquer atividade de construção e pesquisa que envolva corpora de língua natural. Assim, como elucida Costa, *et al*. (2022), é a natureza dos documentos que compõem o corpus que determina o tipo de corpus que se está construindo.

Nesse sentido, os corpora históricos, por exemplo, exigem técnicas e rigor diferentes daqueles utilizados na criação de corpora baseados em textos nativos-digitais, oriundos da web. Conforme Costa, Santos e Namiuti (2021), na iniciativa de construção de corpora históricos baseada no workflow Lapelinc, há uma preocupação, indispensável por sinal, na criação desse tipo de corpus, de uma garantia de fidedignidade e consistência filológica. Essa preocupação gera o desafio de trabalhar com um documento que é físico de uma maneira que seja possível realizar o processamento digital automático, mas sem perder as especificidades e o caráter primordiais do mesmo.

Sendo assim, nessa iniciativa, de acordo com os autores, o Documento Físico (DF) dá suporte para que seja construído um Documento Digital Imagem (DDI). Esse DDI será utilizado no meio digital como fonte original que auxiliará na construção de corpora eletrônicos anotados, integrado por arquivos do tipo Documento Digital Texto (DDT). Nessa transposição de DF para DDI, o método LAPELINC apresenta um Aparato de Metadados Estruturados (AME), composto por cinco componentes básicos que são capazes de garantir, no DDI, a recuperação de diversas informações e complexidades do DF, a saber: i) Catálogo

Visual (CV); ii) Dossiê de Observações Pertinentes (DOP); iii) Fotografia Cientificamente Controlada (FCC); iv) Análise Topográfica (AT) e v) Análise Descritiva (AD). Depois, o documento é tratado pelo método LAPELINC numa etapa de transcrição paleográfica e, a partir disso, se tem um formato de texto puro.

Após as etapas de transposição e transcrição, o texto é transferido para o formato digital de texto simples (TXT), para então ser editado e anotado morfossintaticamente e sintaticamente, sendo portado para o formato XML que é responsável por representar os diferentes módulos de anotação em um único arquivo. Assim, obtém-se finalmente o corpus eletrônico anotado do workflow LAPELINC.

Frente a esses apontamentos, entende-se que, como descrito, o LAPELINC Framework foi desenvolvido para instruir trabalhos de construção de corpora digitais para línguas naturais, para atuar como uma referência padrão para essas línguas. Sendo assim, ratificamos que essa metodologia é viável em fluxos de trabalho de construção de corpora digitais para línguas de sinais, uma vez que essa se configura, também, como língua natural. É pensando nisso que explanaremos, a seguir, um workflow, baseado no LAPELINC Framework, para construção de corpora digitais para línguas de sinais, utilizando especificamente a Libras como exemplo.

5 APRESENTAÇÃO DA PROPOSTA DE WORKFLOW PARA CONSTRUÇÃO DE CORPORA PARA LÍNGUAS DE SINAIS

A partir do levantamento de exemplos de corpora de línguas orais, no que se refere a seus sistemas de anotação, apresentados na seção anterior, tivemos um direcionamento metodológico para a construção de corpora de línguas de sinais, na medida em que, como foi explicitado, já existe para aqueles uma estrutura, um esquema, um workflow para construção de corpora.

Uma vez que um dos nossos objetivos principais foi propor um fluxo de trabalho para construção de corpora de línguas de sinais que atenda as mesmas diretrizes dos corpora orais e escritos no que se refere às possibilidades de anotação e que faça reuso das tecnologias já existentes, nessa seção apresentaremos nossa proposta de workflow a partir da adaptação do framework do método LAPELINC (Costa; Santos; Namiuti, 2021), e experimentaremos ferramentas e tecnologias que já estão sendo utilizadas na construção de corpora de línguas orais e de sinais para a construção do nosso modelo de corpus de língua de sinais, exemplificando a possibilidade de anotação multicamada para corpora de línguas de sinais com um módulo de transcrição da articulação do sinal integrado na primeira camada de anotação linguística com o auxílio da ferramenta ELAN e comparar resultados, atendendo assim aos objetivos específicos propostos na pesquisa.

5.1 O Lapelinc Framework como possibilidade estrutural para a elaboração de um projeto de corpus para línguas de sinais

As etapas para construção de um corpus, baseado no LAPELINC Framework, são, como vimos: (1) Projeto do corpus; (2) Catalogação; (3) Aquisição de mídia; (4) Topografia; (5) Transcrição; (6) Anotação; (7) Recuperação da Informação e (8) Intercâmbio da Informação. Em decorrência disso, agora, nos atemos a elucidar acerca de cada uma delas adaptadas à nossa proposta de workflow de construção de corpus de Libras.

Primeiramente, é preciso ratificar que o nosso projeto se caracteriza como uma proposta de construção de corpora para línguas de sinais, que tem a Libras como exemplo de língua, o qual deve ser composto por um arquivo de vídeo com dados em Libras, um módulo de transcrição e outro de tradução. Sendo assim, a partir da consulta a esse produto de corpus, os pesquisadores serão capazes de realizar consulta de âmbito linguístico, acrescentando

informações como camadas de anotação de estrutura gramatical e fazer inferências, isto é, levantar hipóteses acerca dessa língua.

Acerca da Aquisição de mídia, selecionamos dados de uma pesquisa já existente. Sendo assim, serão utilizados como fontes para guiar a construção do nosso corpus cru (sem anotação) os dados levantados por Ediélia Lavras dos Santos Santana (2019) na sua pesquisa de Mestrado em Linguística do Programa de Pós-Graduação em Linguística – PPGLin, na qual ela investigou sobre a questão da categorização morfológica para nome e verbo em Libras. Para termos acesso aos dados, entramos em contato com a pesquisadora, pedimos permissão e, sendo assim, envio do material.[21]

Explicamos, sobre questões relacionadas ao método, que, pensando em um workflow para construção de corpora, podemos utilizar fontes de três tipos: de corpus já existente e reconfigurá-lo (que foi o caso, como vimos); de fontes digitais disponíveis na internet sem necessariamente já estar em um formato de corpus; ou de fontes construídas do zero que envolve a necessidade, por exemplo, de uma seleção de indivíduos, seleção de informações que se pretende atingir com as fontes e a gravação dos dados. Cada possibilidade dessa dispõe de requisitos importantes para controlar e anotar os metadados na produção de corpora de línguas de sinais, o que engloba a etapa de Catalogação. Nesse sentido, para catalogar os metadados da fonte que utilizamos foi preciso consultar a dissertação de Santana (2019), pois se tratavam de fontes de corpus já existente e, sendo assim, precisávamos conhecer os procedimentos adotados pela pesquisadora. A seguir, expomos as informações pertinentes extraídas de Santana (2019) acerca da construção de seu corpus.

Ela selecionou como sujeitos-informantes 19 pessoas, sendo 4 surdos que compõem o Grupo de Aquisição na Infância (GI), 6 surdos que compõem o Grupo de Aquisição Pós-Infância (GPI) e 9 ouvintes falantes de Libras como segunda língua-L2, que compõem o Grupo Ouvintes (GO). Após diálogo de esclarecimento da pesquisa com os sujeitos-informantes e consentimento por parte dos mesmos, a pesquisadora realizou um levantamento de informações pessoais sobre eles, a fim de constituir o perfil de cada um deles. Santana (2020) esclarece que "Tal detalhamento dos perfis de sujeitos-informantes é importante para compreendermos a que espécie de input estes sujeitos foram expostos e as implicações que recai sobre cada grupo de aquisição" (Santana, 2019, p. 61).

Para a geração dos dados para o corpus, a autora realizou um teste de eliciação de nome e verbo, a partir de imagens em contexto, isto é, eram apresentadas aos sujeitos-

[21] Esses dados são os mesmos que compõem o corpus analisado por nós na subseção 2.2. Esse foi um feito proposital, justamente para termos a possibilidade de, posteriormente, comparar resultados.

informantes imagens contextualizadas das quais fosse possível ter como resultado a obtenção de frases construídas por eles que continham nome e verbo. Exemplo das imagens utilizadas são as a seguir:

Figura 36 – Imagens usadas no teste para geração dos dados


Fonte: Santana, 2019, p. 61.

A primeira imagem foi utilizada para eliciar o verbo PASSAR roupa, na qual encontra-se uma mulher passando um amontoado de roupas. Já a segunda imagem foi selecionada para eliciar o nome FERRO, na qual foi escolhida uma imagem em que o ferro estivesse totalmente destituído de sua função primordial. Antes da aplicação do teste, de fato, foi realizado um teste piloto com um informante de cada grupo para observar se as imagens estavam claras a ponto de ajudarem a levar à obtenção do que foi planejado. O resultado foi positivo, apontando apenas poucos ajustes em algumas imagens.

Ao todo, foram selecionadas 74 imagens para eliciar nomes e verbos para o teste de geração dos dados para a pesquisa. Essas imagens foram dispostas em slide para que fossem projetadas no notebook ou na televisão e organizadas de maneira aleatória, sem sequência com seus respectivos pares, para que não ocorresse nenhum tipo de influência no sinal no momento da gravação das frases. As gravações ocorreram de maneira individual, ou seja, um informante por vez, e contou também com a presença da aplicadora e de um técnico que ficou responsável por manusear o equipamento de filmagem. As sessões se deram algumas na Universidade do Estado da Bahia - UNEB, outras na casa dos sujeitos e outras na casa da pesquisadora; a escolha era feita a partir do que fosse mais viável para cada colaborador.

A pesquisadora afirma que com as gravações dos 19 sujeitos-informantes foi possível alcançar um corpus constituído por conjunto de 1.406 frases produzidas em contexto, envolvendo 25 pares de nomes e verbos: BICICLETA/ANDAR DE BICICLETA[22]; PORTA/

[22] Santana (2019) traz em sua pesquisa o verbo “bicicletar”. Entretanto, substituímos por “andar de bicicleta”, uma vez que estamos, nesse momento, realizando a tradução do sinal e não uma análise por glosa como a pesquisadora. Essa alteração foi necessária pelo fato do verbo “bicicletar” não existir formalmente na língua portuguesa, logo não haver possibilidade de tradução; estudos apontam que há

ABRIRporta; TESOURA/CORTAR COM A TESOURA; PENTE/PENTEAR; CORRIDA/CORRER; NEVE/NEVAR; CHUVA/CHOVER; PENSAMENTO/PENSAR; SONHO/SONHAR; FUTEBOL/CHUTAR; LADRÃO/ROUBAR; EXPLOSÃO/EXPLODIR; NATAÇÃO/NADAR; CONSTRUÇÃO/CONSTRUIR; BRINQUEDO/BRINCAR; BEBIDA/BEBER; COMIDA/COMER; CADEIRA/SENTAR; VENTO/VENTAR; TELEFONE/TELEFONAR; CARRO/DIRIGIRcarro; FERRO/PASSARroupa; CHORO/CHORAR; SORRISO/SORRIR; CASAMENTO/CASAR.

Para finalizar a exposição dessa etapa elaboramos um quadro com os dados que fazem parte do esquema de anotação dessa fase de captura e catalogação. O sistema de anotação desses metadados pode ser feito em XML-TEI, seguindo as recomendações do Lapelinc Framework, associando as tags adequadas para a anotação desses valores. Segue exposto abaixo.

Quadro 22 – Dados que compõem o esquema de anotação

1. Origem do corpus/documento
2. Local onde foi produzida as gravações
3. Data da produção
4. Identificação do informante
5. Idade do informante no momento da captura do vídeo/documento

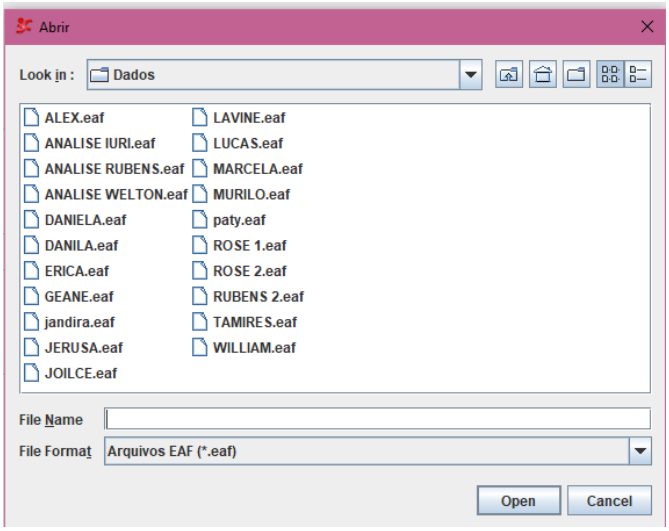
Fonte: Elaboração própria.

Vale a pena ressaltar que as determinações das etapas do workflow discutidas até aqui, atendem, inclusive, aos objetivos (ii) e (iii) de nossa pesquisa, que visam definir os metadados que são importantes para o controle das informações para a construção do corpus e localizar as fontes de dados para a construção do corpus,

O corpus elaborado por Santana (2019) possui dados de 21 informantes, sendo eles: Alex, Daniela, Danila, Érica, Geane, Iuri, Jandira, Jerusa, Joilce, Lavine, Lucas, Marcela, Murilo, Paty, Rose 1, Rose 2, Rubens 1, Rubens 2, Tamires, Welton e William, conforme apresentados abaixo.

uma chance de existir em Libras devido à forma como é executado o sinal. (o mesmo acontece com o verbo “tesourar” quando substituímos por “andar de bicicleta”)

Figura 37 – Dados de Santana (2019)



Fonte: Print Screen da pasta dos arquivos no computador.

Ao clicar nas pastas, é apresentado os dados em forma de frases gerados a partir do contato com cada informante. Utilizaremos os dados da informante Jerusa. Em sua base de dados, encontramos 57 frases traduzidas por glosa, as quais listamos no quadro a seguir:

Quadro 23 – Frases realizadas pela informante Jerusa

01	ESTE HOMEM GOSTAR CORR[er]/CORR[ida] 4X
02	HOMEM QUER IR CASA PESSOA BATER (porta) não incorporado (argumento nulo)
03	HOMEM SENTAR/CADEIRA PENS[ar]/PENS[amento]
04	MULHER HOMEM OS DOIS CAS[ar]/CAS[amento]^ ASSINAR uso do composto
05	CONSTR[uir]/CONSTR[ução] AGORA PRÉDIO CONCLUIR 4x
06	PESSOAS MENINA SENTAR/CADEIRA LER TER SENTAR/CADEIRA
07	HOMEM CHEGAR ATÉ PORTA/ABRIRporta
08	EQUIPE TEM CORR[er]/CORR[ida] UM PARTICIPANTE CHEGAR VENCER 2x
09	TRES AMIGOS PASSEAR BICICLETA/ANDARbicicleta 4x
10	EQUIPE FUTEBOL/CHUTAR COMEMORAMCM 1
11	EQUIPE PESSOAS NADAR/NATAÇÃO 6X
12	HOMEM O CEREBRO PENS[ar]/PENS[amento] CM D e 4
13	---------
14	MULHER BEB[er]/BEB[ida] AGUA CM C
15	HOMEM BICICLETA/ANADARbicicleta VER LIMPOcl 2x
16	HOMEM COMEÇOU IRRITADO SENTAR/CADEIRA EMPURRAR CAIR
17	MULHER GUARDA-CHUVA VENT[o]/VENT[ar]forte LEVANDO 2x
18	CRIANÇA BRINC[ar]/BRINQ[uedo] MONTAR PEÇAS BOM AMIGO 4 x curto
19	HOMEM CONSTR[uir]/CONSTR[ução]^ CM S e B 2X e 6x
20	PESSOA FAZER CONVITE DAR CAS[ar]/CAS[amento]
21	MULHER PRECISA FERRO/PASSAR-ROUPA^ROUPA ARRUMAR GUARDAR 3x composto
22	MULHER FOI SUPERMERCADO COMPRAR VARIOS BEB[er]/BEB[ida]^ REFRIGERANTE, CERVEJA, CACHAÇA, COMPRAR VARIOS CASA detalhou as bebidas
23	MULHER DORMIR DESEJOSA SONH[o]/SONH[ar] VONTADE COM[er]/COM[ida] BOLO
24	*******
25	BRINC[ar]/BRINQ[uedos] VARIOS CACHORRO composto varios

26	*********
27	MENINO Pens[AR]/PENS[amento] CHOR[o]/CHOR[ar] SOFRER CHOR[o]/CHOR[ar]muito
28	TER RUA CASA TER CHUV[a]/CHOV[er] FORTEmuito
29	******
30	MULHER FOME PRECISA COM[er]/COM[ida]
31	TEM TRÊS TELEF[one]/TELEF[onar]. UM ANTIGO, UM ATUAL NOVO TELEF[one]/TELEF[onar]
32	HOMEM LER CARTA EMOCIONA CHOR[o]/CHOR[ar]
33	MULHER OLHAR CABELO QUERER TESOURA/CORTARtesoura
34	GRUPO HOMENS VARIOS ROUBAR/LADRÃO FORAM PRESOS
35	TEM CHUV[a]/CHOV[er] FORTEmuito TER AGUA ALAGAR
36	HOMEM QUER LIMPO CARRO/DIRIGIRcarro LAVARcl
37	TER COM[ida]/COM[er] REJEITAR JOGAR LIXO
38	PESSOA CARRO/DIRIGIRcarro TER BEB[er]/BEB[ida] NÃO-PODE PERIGO
39	****** MULHER SABER NADAR/NATAÇÃO BEM
40	TEM TESOURA/CORTARtesoura TATUAGEMbraço
41	MULHER HOMEM CAS[ar]/CAS[amento] BRINC[ar]/BRIN[quedos] NEVE/NEVAR DIVERTIR
42	****PASSAR-ROUPA/FERROTEM CARNE LINGUIÇA FRITAR COM[er]/COM[ida]
43	PESSOA PENT[e]/PENT[ear] QUEBRAR
44	*****
45	ER ANDANDO BOLSA HOMEM VONTADE PEGAR CEULAR ROUBAR/LADRÃO
46	HOMEM TEM SORR[ir]/SORR[iso]**** DIFERENTE
47	MULHER PENT[e]/PENT[ear] ****
48	TEM EXPLO[são]/EXPL[odir] pequenas EXPLO[são]/EXPL[odir] grande
49	MULHER TELEF[one]/TELEF[onar] SUPORTARCABEÇA DOER CANSADA
50	ARVORES TER VENT[o]/VENT[ar]forte ARVORES BALANÇAR ****
51	HOMEM QUERER DINAMITE JOGAR BANCO EXPL[odir]/EXPLOS[são] QUER ROUBAR BANCO
52	*** MULHER SONH[o]/SONH[ar] PESSOA ASSOMBRAR MEDO
53	GRUPO AMIGOS GOSTAR FUTEBOL/CHUTAR/ ESPORTES ANDARbicicleta, CORR[er]/CORR[ida], NADAR/NATAÇÃO
54	PESSOA ROUBAR/LADRÃO CHEGAR CASA ABRIR-PORTA/PORTA PROCURAR VARIOS ROUBAR/LADRÃO
55	MENINO QUERER VONTADE ENTRAR MAS ABRIR-PORTA/PORTA DIFICULDADE
56	CEREBRO PENS[ar]/PENS[amento] oque? FUTURO O QUE? QUER TRABALHAR FACULDADE VARIOS
57	EXEMPLO CAIR BOMBA EXPLODIR/EXPLOSÃO ENORME (ocorrendo como argumento interno do existencial ter)

Fonte: Elaboração própria a partir dos dados de Santana (2019).

Essas informações sumarizadas, alcançadas a partir da divisão do vídeo em frases com identificação, constituem a etapa de Topografia do corpus.

Por sua vez, a etapa de Transcrição da nossa iniciativa será constituída pelo Sistema de Escrita da Libras, a escrita Sel, isto é, a Libras será transcrita para a Sel. O módulo de transcrição tem a finalidade de permitir uma associação entre forma e sentido por parte do linguista e/ou de uma ferramenta computacional, proporcionando, por exemplo, a formulação de hipóteses sobre a língua. Posto isso, esclarecemos que elegemos a escrita Sel para integrar esse módulo porque, como vimos na seção 3, ela atende satisfatoriamente a esse objetivo, uma vez que se mostrou ser uma escrita bem fundamentada e eficiente, que representa com precisão a estrutura fonológica do sinal. Portanto, com a Sel, um pesquisador, ao olhar para a

transcrição presente em nossa proposta, tendo como pré-requisito o conhecimento das regras de funcionamento dessa escrita, será capaz de realizar o sinal em Libras, alcançando, assim, a associação entre forma (transcrição) e sentido (sinal).

É preciso elucidar que com o que foi mencionado sobre a escrita Sel e com essa etapa de Transcrição, atendemos ao objetivo (i) de nossa pesquisa, sendo ele o de verificar se a escrita Sel poderá atender aos requisitos tecnológicos da etapa de transcrição.

Na etapa de Anotação, se dará a tradução da Libras por meio da língua portuguesa. Nos corpora de línguas de sinais, a tradução é feita por meio de uma língua estabelecida pelos pesquisadores que construíram cada corpus; desse modo, no nosso caso, achamos viável utilizar a língua portuguesa por se tratar da nossa língua materna, a primeira língua aprendida por nós. Além dessa anotação por meio de tradução, futuramente, será pensada também uma anotação referente à gramática da Libras (morfológica, sintática e semântica). Esse será um trabalho posterior, pois esse tipo de anotação para a Libras demanda mais tempo e dedicação porque como ela é uma língua ainda pouco estudada no que diz respeito, principalmente, a sua estrutura gramatical, torna-se difícil encontrar pesquisas que discutam e ratifiquem informações, por exemplo, do que é classe gramatical (substantivo, verbo, adjetivo) ou função sintática (sujeito, predicado, aposto) na Libras.

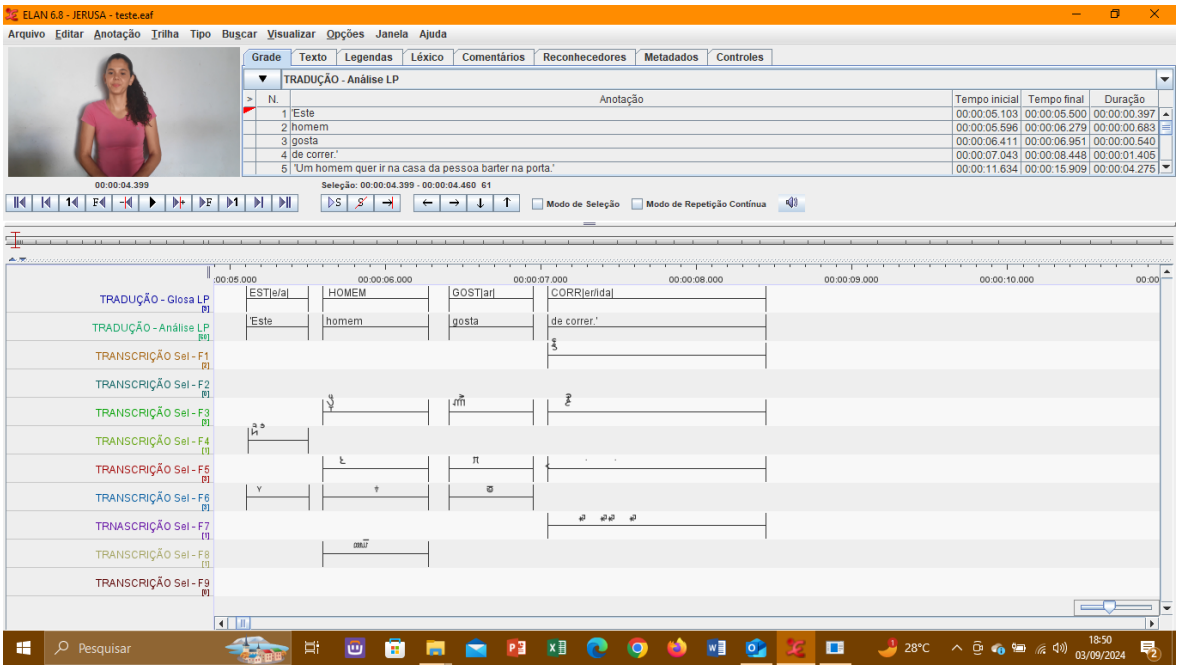
Toda essa estrutura, em sentido de registro de informações, mencionada até aqui dá condições para que as duas últimas etapas existam. Desse modo, as etapas de Recuperação da Informação – que é responsável por buscas automáticas, por exemplo – e Intercâmbio da Informação – que é a possibilidade de importar/exportar textos do corpus – são possíveis em nossa iniciativa, entretanto só serão capazes de serem realizadas quando uma ferramenta apropriada for desenvolvida, pois dependem de uma tecnologia que o ELAN não oferece. Um adendo sobre a etapa de Recuperação é que ela pode ser também manual, porém é interessante que ela seja automática, pois se tratando de um corpus extenso e com muitas informações, isso contribui com a otimização do tempo e do trabalho, além de demonstrar ser uma iniciativa tecnológica mais avançada.

Posto isso, esclarecemos que o ELAN ainda não é a ferramenta ideal, uma vez que não é livre, possui uma tecnologia rudimentar e ainda precisa do uso do editor SEL, mas por hora foi utilizada para proporcionar o alinhamento entre o vídeo e a transcrição. Em uma pesquisa futura, será pensada uma ferramenta que seja aberta e que utilize de padrões de anotação que são conhecidos e utilizados no mundo e que essa anotação possa ser transferida para qualquer outra ferramenta. Ela foi utilizada apenas para projetar o workflow, para em pesquisa futura serem desenvolvidas ferramentas que atendam de fato ao projeto.

5.2 A possibilidade de anotação multicamada e reuso de tecnologia para um modelo de construção de corpora de língua de sinais

Será apresentada, a seguir, um experimento de reuso de tecnologias para aplicação de um modelo de construção de corpora multicamada para línguas de sinais e, em seguida, será realizada uma comparação desse com os corpora construídos com a utilização de outros esquemas. Esclarecemos que com isso atendemos aos objetivos (iv) e (v) da pesquisa, que são, respectivamente, exemplificar a aplicação do workflow proposto para a construção de corpora de língua de sinais em dados de Libras extraídos de dissertações e teses consultadas na pesquisa, com o objetivo de testá-lo e comparar os resultados obtidos com outras iniciativas de construção de corpora para língua de sinais. A figura 38 apresenta a tela da ferramenta ELAN escolhida para o teste com linhas de transcrição utilizando a escrita SEL, tecnologia escolhida para a camada de transcrição do corpus.

Figura 38 – Amostra do experimento de construção de corpora para línguas de sinais



Fonte: Print Screen da tela do ELAN

Esse produto do experimento é composto por um arquivo de vídeo com dados em Libras o qual se encontra exposto no canto esquerdo da tela e, logo abaixo do arquivo, temos trilhas de segmentos organizados da seguinte maneira: uma trilha com informações de tradução por glosa da língua portuguesa, uma trilha de tradução com análise da língua

portuguesa[23] e, em seguida, nove trilhas configuradas de F1 a F9 com informações do módulo de transcrição por escrita Sel. É importante ratificar que como com esse módulo de transcrição por escrita Sel se tem o sinal anotado, o pesquisador é capaz de olhar para a anotação e realizar o sinal gesto-visual, tendo como pré-requisito o conhecimento dos mecanismos dessa escrita.

Todas essas informações referentes à anotação estão interligadas e, ao dar play no arquivo de vídeo, o marcador faz a relação entre o sinal que está sendo realizado e as anotações referentes a esse sinal. Sem dúvidas, essa possibilidade de fazer uma relação do sinal com a anotação contribui com a análise que será feita pelo pesquisador. Inclusive, com essa iniciativa de construção de corpora para línguas de sinais, é possível criar outras trilhas com informações gramaticais de caráter morfológico, sintático e semântico, caso seja o objetivo da pesquisa.

Aqui é importante ressaltar que as tecnologias escolhidas (a ferramenta ELAN e a escrita SEL) impuseram um desafio ao experimento e demonstraram um limite em relação à sua utilização em conjunto. Destacamos que o estado atual da escrita SEL em meio eletrônico ainda é rudimentar, não existe uma fonte única que escreva seus caracteres, o sistema de escrita no computador necessita de combinar diversas fontes o que traz uma complexidade computacional que não pode ser ignorada e uma dificuldade para o alinhamento utilizando a ferramenta ELAN. Como é possível observar na figura 38, tivemos que fazer uma anotação fragmentada: para transcrever um sinal foram necessárias 9 linhas de transcrição. Essa dificuldade se deu devido ao fato da escrita Sel ser composta, como vimos na seção 3.3.4, por nove fontes de caracteres que atuam concomitantemente para o funcionamento da escrita – na escrita Sel do sinal QU|er|  , por exemplo, são utilizados o teclado F3 de configurações da mão direita minúsculas e o teclado F6 de movimentos de mão. Todavia, no ELAN, só é aceita uma fonte por trilha, daí a necessidade de construir nove trilhas para atender às necessidades da escrita Sel.

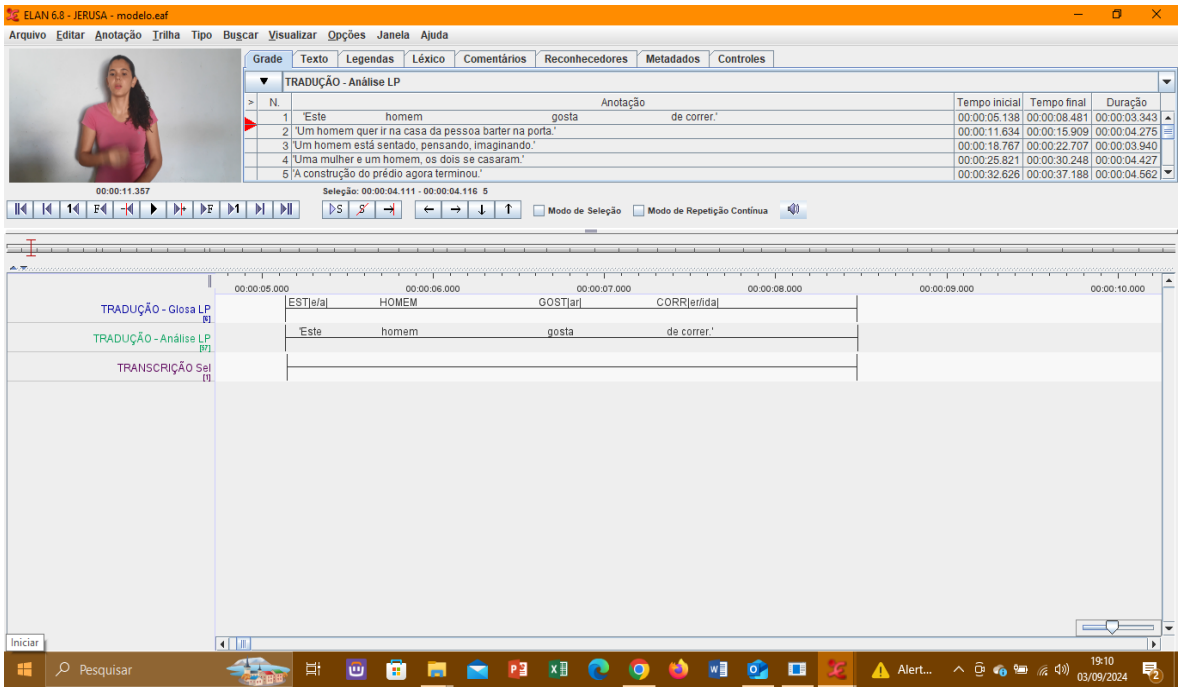
Este fato trouxe um limite importante para a utilização desta escrita associada a ferramenta ELAN, pois não foi possível anotar a transcrição por Sel em uma trilha apenas, como é feito com as línguas orais através do alfabeto fonético na fonte IPA.

[23] Adicionamos dois módulos de tradução – um por glosa e outro por análise – porque a tradução por glosa está livre de inferências como no excerto “CORR|er/ida|”, diferentemente da tradução por análise, na qual já se afirma que este excerto se realiza como um verbo no infinitivo acompanhado de uma preposição “de correr”. Sendo assim, para sermos mais responsáveis com os dados, achamos viável trazer as duas maneiras de tradução.

Com o experimento concluímos que há uma necessidade de adaptação das tecnologias, tanto da escrita quanto da ferramenta, para que sua utilização seja viável. Uma escrita que necessite de várias fontes traz uma complexidade de processamento computacional importante e indica o limite da ferramenta ELAN na verdade é um limite do estado da arte da tecnologia da escrita SEL, para que ela seja uma opção viável para se integrar a um modelo computacional de construção de corpora que reutiliza tecnologias e ferramentas já desenvolvidas para construção de corpora orais há ainda que se buscar desenhar/construir uma única fonte que consiga escrevê-la digitalmente, integrando-a no padrão Unicode de alfabetos digitais.

Abaixo, apresentamos o modelo teórico de como seria a anotação multicamada utilizando a escrita SEL. Nesse, há apenas uma trilha para cada módulo criado. Assim, torna-se mais fácil a associação e, consequentemente, a análise:

Figura 39 – Projeto de modelo teórico ideal de apresentação de anotações

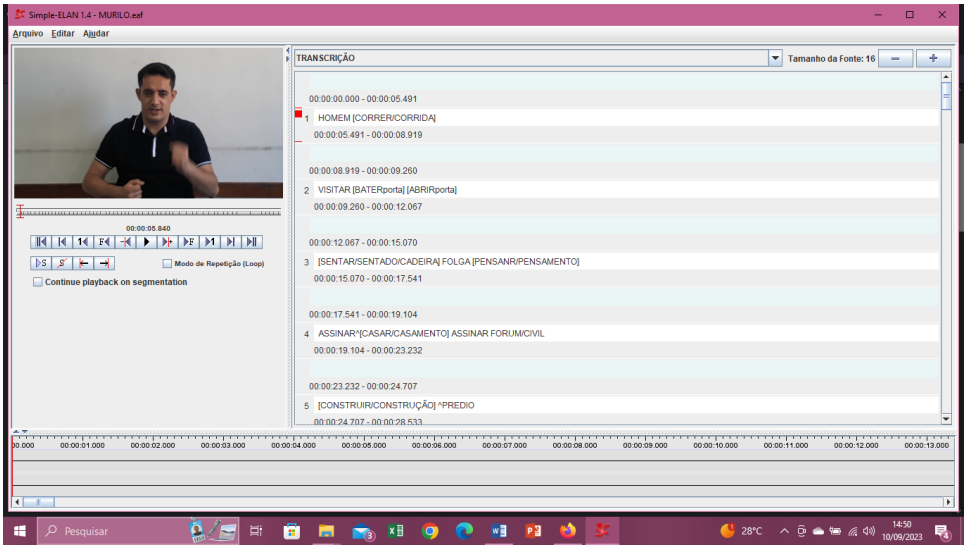


Fonte: Print Screen da tela do ELAN (adaptado por nós com colagem de anotação de transcrição em escrita Sel).

Ratificamos que a amostra da figura 38 é o produto do experimento de construção de corpora para línguas de sinais realizado a partir da ferramenta Elan tal como ela atualmente se encontra, a qual foi a alternativa viável e possível de ser utilizada no momento. No entanto, o nosso modelo teórico, pensado para a iniciativa de construção de corpora para línguas de sinais se configura como o da figura 39.

Comparamos o modelo teórico construído e exemplificado na figura 39 com algumas das iniciativas existentes de construção de corpora e citadas no decorrer do trabalho. Santana (2019) também construiu o seu corpus utilizando a ferramenta ELAN, como podemos verificar abaixo na Figura 40 (reexibição da figura 3).

Figura 40 – Programa ELAN, corpus de Santana (2019)

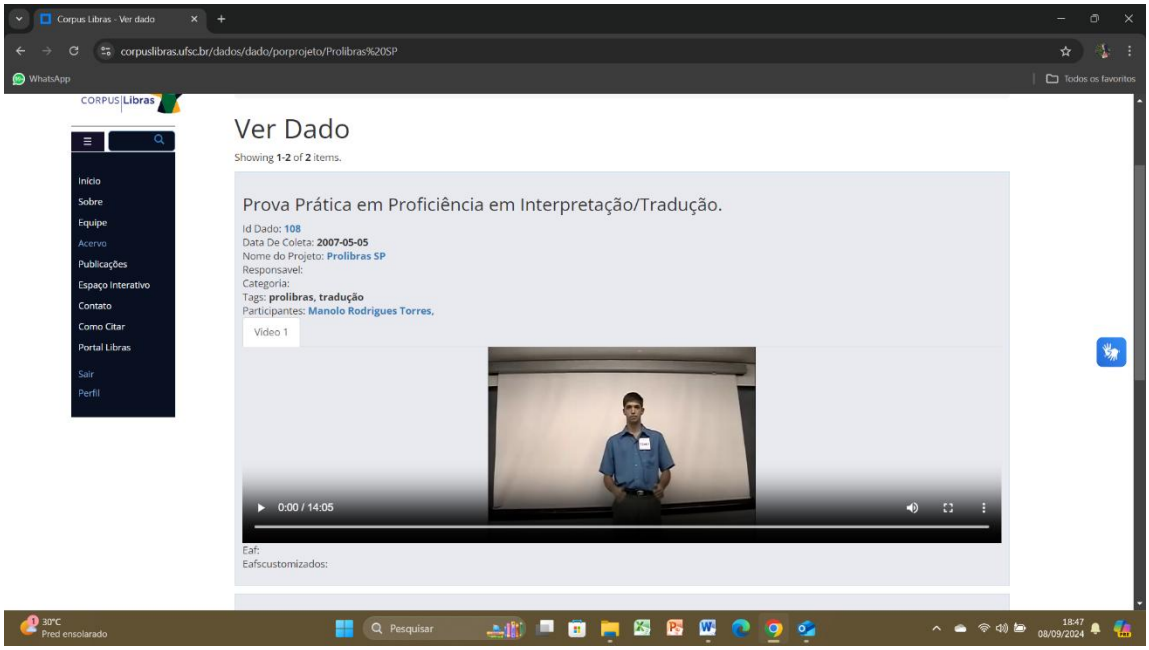


Fonte: Print Screen dos dados do informante Murilo

Esse exemplo, assim como o nosso, foi construído no ELAN, porém possui informações distintas. Nele encontramos um arquivo de vídeo e anotações vinculadas a esse vídeo. O problema é que essas anotações se dão apenas por transcrição por glosa – há uma ausência de tradução para língua portuguesa, por exemplo, e de transcrição fonética-fonológica –, o que, como já discutimos, pode trazer uma limitação para as análises dos pesquisadores.

Trazemos, também, uma segunda iniciativa existente: a do Corpus Libras, exemplificada na figura 41.

Figura 41 – Iniciativa do Corpus Libras



Fonte: Print Screen da tela do site do Corpus Libras

Ao compararmos essa iniciativa com o nosso modelo de anotação, percebemos que nessa se encontra apenas o dado cru com anotação de proveniência (metadados) e sem anotação linguística, como transcrições, glossas ou traduções, ou seja, traz apenas um arquivo de vídeo em Libras e um levantamento de informações a respeito da coleta de dados, como data de coleta e nome do projeto.

Essas comparações demonstram que há uma carência de corpora anotados linguisticamente para o estudo da Libras, o que indica a necessidade de programar e desenvolver projetos futuros com o intuito de realizar uma construção padronizada como a que propusemos no modelo defendido nessa dissertação.

6 CONSIDERAÇÕES FINAIS

Tivemos como um dos nossos objetivos fazer um levantamento sobre a forma que os estudos da área lidam com os dados da língua de sinais, no que se refere à maneira como são construídos os corpora que dão suporte às pesquisas, as possibilidades e os limites encontrados. Verificamos que as iniciativas de construção de corpora de Libras não possuem módulo de transcrição do sinal o que se demonstrou um problema para as investigações na área da Linguística. A partir disso, propomos um fluxo de trabalho para construção de corpora de línguas de sinais, que possa atender as diretrizes dos corpora orais e escritos no que se refere às possibilidades de anotação e que reutilize as tecnologias já existentes.

Portanto, com base nos estudos e resultados obtidos nesse trabalho, foi possível demonstrar a importância que tem um corpus bem estruturado para a pesquisa linguística e, sendo assim, entendemos que novas iniciativas de construção de corpora para línguas de sinais são essenciais, uma vez que as existentes ainda carecem de aprimoramentos para que atendam satisfatoriamente às exigências da área da Linguística de Corpus. Isso porque, como vimos, os corpora de línguas de sinais atuais não seguem orientações metodológicas e padronizadas como é recomendado pela linguística de corpus.

Frente a esses apontamentos, fica evidente que nossa pesquisa, sendo um workflow para construção de corpora em línguas de sinais, contribui com essa questão, pois se trata de um trabalho que tem como resultado uma orientação metodológica para essas iniciativas. Em outras palavras, é uma proposta que ajudará a padronizar a atividade de construção de corpora para línguas sinais e, dessa maneira, contribuir com a solução de problemas existentes como a ausência de um módulo de anotação de transcrição fonética ou fonológica. Também avaliamos as propostas dos sistemas de escrita para língua de sinais e, dentre as escritas apresentadas, a escrita Sel se mostrou a mais viável no momento para exercer essa função da transcrição, devido a sua capacidade de representar com precisão a estrutura fonológica do sinal.

Como trabalhos futuros, será necessário para o desenvolvimento do modelo aqui defendido uma adequação da tecnologia da escrita SEL e das ferramentas existentes para atender aos requisitos computacionais que a anotação de corpora multicamada pressupõe.

Por fim, esta pesquisa é uma inovação para as iniciativas de construção de corpora para línguas de sinais, uma vez que se caracterizou como delineamento de um modelo de construção de corpora para Libras que melhor atendesse às pesquisas em linguística. Com isso, ela tem o potencial de auxiliar o público acadêmico que tem interesse em investigar,

descrever e analisar a Libras, uma vez que essa é uma língua que ainda carece de estudos em todos os âmbitos por ter sido reconhecida muito recentemente como língua natural. Desse modo, toda reunião de esforços é imprescindível para que cada vez mais essa língua se torne reconhecida como de fato língua e que tenha sua estrutura gramatical bem analisada.

REFERÊNCIAS

ALUÍSIO, Sandra Maria; ALMEIDA, Gladis Maria de Barcellos. O que é e como se constrói um corpus?: lições aprendidas na compilação de vários corpora para pesquisa lingüística. **Calidoscópio**, Unisinos, v. 4, n. 3, p. 156-179, dez. 2006.

BARBOSA, Thaís Bolgueroni. **Uma descrição do processo de referenciação em narrativas contadas em língua de sinais brasileira (Libras)**. 2013. 155 f. Dissertação (Mestrado) - Curso de Linguística, Universidade de São Paulo, São Paulo, 2013.

BARROS, Mariângela Estelita. **ELiS – Escrita das Línguas de Sinais**: proposta teórica e verificação prática. 2008. 199 f. Tese (Doutorado) - Curso de Pós-Graduação em Linguística, Universidade Federal de Santa Catarina, Florianópolis/Sc, 2008.

BARROS, Mariângela Estelita. Princípios básicos da ELiS:: escrita das línguas de sinais. **Revista Sinalizar**, São Paulo, v. 1, n. 2, p. 204-210, dez. 2016.

BENASSI, Claudio Alves. Visografia: uma nova proposta de escrita da língua de sinais. **Traços de Linguagem**, Cáceres, v. 2, n. 2, p. 71-82, 2018.

Capovilla, F. C.; RAPHAEL, W.D. (Org.) ; MAURICIO, A.C. (Org.) . **Novo Deit-Libras: Dicionário enciclopédico ilustrado trilíngue da Língua de Sinais Brasileira (Libras) baseado em linguística e neurociências cognitivas**, 2a. edição revista e ampliada, Volume 2: Sinais de I a Z.. 2. ed. São Paulo, SP: Edusp, 2011. v. 1. 2759p.

Capovilla, F. C.; RAPHAEL, W.D. (Org.) ; MAURICIO, A.C. (Org.) . Novo Deit-Libras: Dicionário enciclopédico ilustrado trilíngue da Língua de Sinais Brasileira (Libras) baseado em linguística e neurociências cognitivas, 2a. edição revista e ampliada, Volume 1: Sinais de A a H.. 2. ed. São Paulo, SP: Edusp, 2011. v. 1. 1418p.

COSTA, Aline Silva. **WEBSINC**: uma ferramenta web para buscas sintáticas e morfossintáticas em corpora anotados - estudo de caso do corpus dovic - bahia. 2015. 186 f. Dissertação (Mestrado) - Curso de Linguística, Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2015.

COSTA, B. S. *et al*. The Systematic Construction of Multiple Types of Corpora Through the Lapelinc Framework. In: Vládia Pinheiro; Pablo Gamallo; Raquel Amaro; Carolina Scarton; Fernando Batista; Diego Silva; Catarina Magro; Hugo Pinto (Eds.) **Computational Processing of the Portuguese Language**. Springer, Switzerland, 2022.

COSTA, B. S; SANTOS, Jorge Viana; NAMIUTI, Cristiane. **Uma proposta metodológica para a construção de corpora através de estruturas de trabalho: o Lapelinc Framework.** Revista Brasileira em Humanidades Digitais, [S. l.], v. 1, n. 2, 2021.

COSTA, B. S. **Um framework integrado para a criação, o gerenciamento e a disponibilização de corpora digitais em língua portuguesa**. Projeto de Pesquisa de Doutorado (PPGLIN/UESB). Vitória da Conquista. 2019.

Finger, Marcelo; SOUSA, M. C. P. ; NAMIUTI, C. ; MONTE, V. M.. **Corpus Carolina** v1.0 Ada. 2022. (Corpus).

Finger, M., Paixão de Souza, M. C., Namiuti, C., Monte, V. M., Costa, A. S., Serras, F. R., Sturzeneker, M. L., Guets, R. P., Mesquita, R. M., Crespo, M. C. R. M., Rocha, M. L. S. J., Palma, M. F., Silva, M. M., Brasil, P. **Carolina**: a General Corpus of Contemporary Brazilian Portuguese with Provenance and Typology Information. Language resources and evaluation, submitted paper (2021).

GALVES, Charlotte. O corpus tycho brahe: um corpus sintaticamente anotado do Português histórico. **Rbba**, Vitória da Consquista, v. 1, n. 8, p. 181-204, jul. 2019.

GALVES, Charlotte; SANDALO, Filomena; SENA, Ticiana A. de; VERONESI, Luiz. Annotating a polysynthetic language: from portuguese to kadiwéu. **Cadernos de Estudos Lingüísticos**, [S.L.], v. 59, n. 3, p. 631, 4 dez. 2017. Universidade Estadual de Campinas. http://dx.doi.org/10.20396/cel.v59i3.8651003.

GALVES, Charlotte; SANDALO, Filomena; SENA, Ticiana A. de; VERONESI, Luiz. **Corpus Kadiwéu**. 2017. Disponível em: https://www.tycho.iel.unicamp.br/viewer/C12. Acesso em: 17 mar. 2024.

Grishman, R. (1996). **TIPSTER Text Phase II Architecture Design**. Version 2.1p. Computer Science. New York University.

JEREMIAS, Daiana do Amaral. **Iconicidade nas sentenças topicalizadas da Libras**: uma motivação semântica e pragmática. 2020. 215 f. Tese (Doutorado) - Curso de Linguística, Universidade Federal de Santa Catarina, Florianópolis, 2020.

KADER, Cárla Callegaro Corrêa; RICHTER, Marcos Gustavo. Linguística de corpus: possibilidades e avanços. **Instrumento**, Juiz de Fora, v. 15, n. 1, p. 13-23, jan/jun. 2013.

LESSA-DE-OLIVEIRA, Adriana S. C. **Componentes articulatórios da Libras e a escrita SEL Estudos da Língua(gem)**, Vitória da Conquista, v. 17, n. 2, p. 103-122, 2019.

LESSA-DE-OLIVEIRA, Adriana S. C. **Libras escrita**: o desafio de representar uma língua tridimensional por um sistema de escrita linear. ReVEL, v. 10, n. 19, 2012.

LESSA-DE-OLIVEIRA, Adriana S. C. **Por uma modalidade escrita da Libras**: estrutura frasal e sinalização, a estrutura fonológica do sinal e a escrita sel. Campinas, Sp: Pontes Editores, 2023. 179 p.

MAGRO, Catarina; VAAMONDE, Gael. Atlas sintático do Português europeu. **Revista Binacional Brasil-Argentina**: Diálogo entre as ciências, [S.L.], v. 8, n. 1, p. 249, 31 jul. 2019. Universidade Estadual do Sudoeste da Bahia/Edicoes UESB. http://dx.doi.org/10.22481/rbba.v8i1.5593.

MCCARTHY, M.; O'KEEFFE, A. Historical perspective: what are corpora and how have they evolved? In OKEEFFE, A.; McCARTHY, M. **The Routledge Handbook of Corpus Linguistics**. New York: Routledge, 2010.

McENERY, T.; WILSON, A. **A corpus linguistics**. Edinburg: Edinburg University Press, 1997.

MENGEL, Andreas; LEZIUS, Wolfgang. An XML-based representation format for syntactically annotated corpora. **International Conference On Language Resources And Evaluation**, Stuttgart, maio 2000.

MONTEIRO, Myrna Salerno. **A interferência do Português na análise gramatical em Libras**: o caso das preposições. 2015. 250 f. Dissertação (Mestrado) - Curso de Linguística, Universidade Federal de Santa Catarina, Florianópolis, 2015.

MOREIRA, Daniele Santana; ROSADO, Luiz Alexandre da Silva. A importância da escrita das línguas de sinais: mapeando propostas e resultados de aplicação na literatura acadêmica nacional. **Revista Espaço**, Rio de Janeiro, v. 54, n. 1, p. 187-208, dez. 2020.

MOREIRA, Renata Lúcia. **Um Olhar da Semiótica para os Discursos em Libras**: descrição do tempo. 2016. 207 f. Tese (Doutorado) - Curso de Linguística, Universidade de São Paulo, São Paulo, 2016.

OTHERO, G.A. Linguística Computacional: Uma breve introdução. Letras de Hoje, Porto Alegre v.41, n.2, 2006.

PRADO, Lizandra Caires do. **Sintaxe dos determinantes na língua brasileira de sinais e aspectos de sua aquisição**. 2014. 163 f. Dissertação (Mestrado) - Curso de Linguística, Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2014.

QUADROS, Ronice Müller de. A transcrição de textos do Corpus de Libras. **Revista Leitura**: Línguas de Sinais: abordagens teóricas e aplicadas, São Paulo, v. 1, n. 57, p. 8-34, jun. 2016.

QUADROS, Ronice M. de.; SCHMITT, Deonísio; LOHN, Juliana T.; LEITE, Tarcísio de A. **Corpus de Libras**. http://corpuslibras.ufsc.br/ 2020.

ROCHA, Amanda Oliveira. **Uma investigação sobre o uso de recursividade em Libras**. 2021. 133 f. Dissertação (Mestrado) - Curso de Linguística, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021.

SAMPAIO, Adovaldo Fernandes. Breve história da escrita. In: SAMPAIO, Adovaldo Fernandes. **Letras e Memória**: uma breve história da escrita. São Paulo: Ateliê Editorial, 2009. p. 13-293.

SAMPAIO, Thamires Oliveira de Souza. **A natureza gramatical da Libras adquirida por surdos e ouvintes**: sinal, classificador, ação construída e gesto. 2020. 169 f. Dissertação (Mestrado) - Curso de Linguística, Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2020.

Sampson, Geoffrey. English for the Computer: The SUSANNE Corpus and Analytic Scheme. Clarendon Press, 1995.

SANTANA, Ediélia Lavras dos Santos. **A questão da categorização morfológica para nome e verbo em Libras**. 2019. 140 f. Dissertação (Mestrado) - Curso de Linguística, Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2019.

SANTORINI, B. **Annotation manual for the Penn Historical Corpora and the PCEEC**. Disponível em: https://www.ling.upenn.edu/hist-corpora/annotation/index.htm. 2010. Acesso em: 02 mar. 2024.

SANTOS, Jorge Viana; NAMIUTI, Cristiane. O futuro das Humanidades Digitais é o passado. In: CARRILHO, Ernestina; MARTINS, Ana Maria; PEREIRA, Sandra; SILVESTRE, João Paulo (org.). **Estudos Linguísticos e Filológicos Oferecidos a Ivo Castro**. [S.L.]: Centro de Linguística da Universidade de Lisboa, 2019. p. 1381-1403.

SARDINHA, Tony Berber. **Linguística de Corpus**. Barueri: Manole, 2004.

SIGNPUDDLE - Online. Disponível em: https://www.signbank.org/signpuddle2.0/signmaker.php?ui=12&sgn=46. Acesso em: 27 nov. 2022.

SILVA, Alan David Sousa; COSTA, Edivaldo da Silva; BÓZOLI, Daniele Miki Fujikawa; GUMIERO, Daniela Gomes. OS SISTEMAS DE ESCRITA DE SINAIS NO BRASIL. **Revista Virtual de Cultura Surda**, Rio de Janeiro, v. 23, p. 1-30, maio 2018.

SILVA, Anderson Almeida da. **A (in)definitude no sintagma nominal em Libras [recurso eletrônico]**: uma investigação na interface sintaxe-semântica. 2019. 351 f. Tese (Doutorado) - Curso de Linguística, Universidade Estadual de Campinas, Campinas, 2019.

SILVA, F. I. da. Ler em SignWriting: possibilidades de desenvolvimento cognitivo da criança surda. In: PERLIN, G.; STUMPF, M. (Orgs.). **Um olhar sobre nós surdos: leituras contemporâneas**. Curitiba: CRV, 2012, p.199-211

SILVA, Igor Valdeci Ramos da. **Aspectos de nomes e verbos na Libras**: identificação morfossintática. 2020. 157 f. Dissertação (Mestrado) - Curso de Linguística, Universidade Federal de Santa Catarina, Florianópolis, 2020.

SILVA, Ione Barbosa de Oliveira. **A categoria dos verbos na língua brasileira de sinais**. 2015. 174 f. Dissertação (Mestrado) - Curso de Linguística, Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2015.

SILVEIRA, F.P. **Integração de ferramentas para compilação e exploração de corpora**. 2008. 101 f. Dissertação (Mestrado em Ciência da Computação) - Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2008.

Skut, Wojciech; Brants, Thorsten; Krenn, Brigitte; Uszkoreit, Hans. A Linguistically Interpreted Corpus of German Newspaper Text. Workshop on Recent Advances in Corpus Annotation, 1998.

Sturzeneker, Mariana Lourenço; Crespo, Maria Clara Ramos Morales; Rocha, Maria Lina de Souza Jeannine; Finger, Marcelo; Paixão de Sousa, Maria Clara; Monte, Vanessa Martins do; Namiuti, Cristiane. **'Carolina's Methodology**: building a large corpus with provenance and typology information'. Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022). CEUR-WS, Vol. 3128, 2022. Available at http://ceur-ws.org/Vol-3128.

VELOSO, Brenda Silva. **Construções classificadoras e verbos de deslocamento, existência e localização na língua de sinais brasileira**. 2008. 159 f. Tese (Doutorado) - Curso de Linguística, Universidade Estadual de Campinas, Campinas, 2008.

VILAÇA, M.L.C. Pesquisa e ensino: Considerações e reflexões. **Revista e-scrita**. Uniabeu, v.1, n.2, 2010.

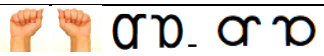
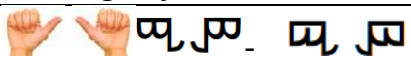
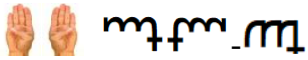
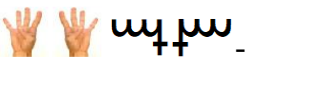
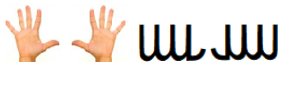
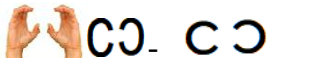
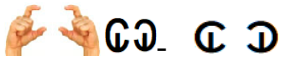
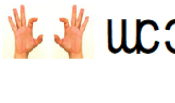
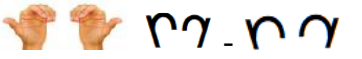
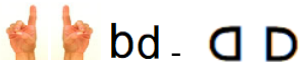
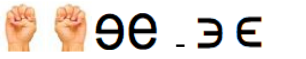
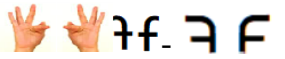
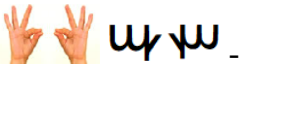
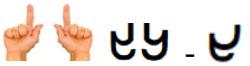
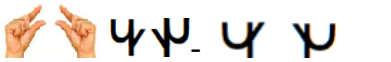
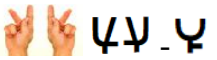
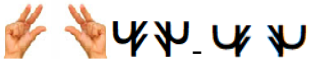
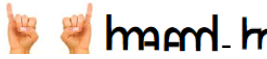
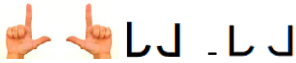
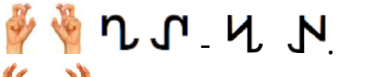
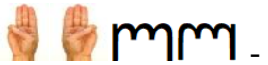
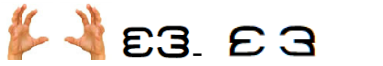
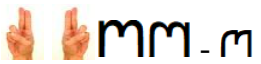
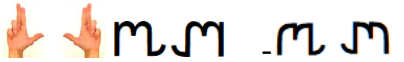
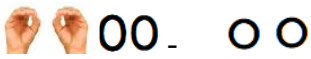
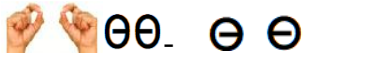
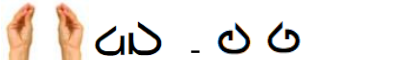
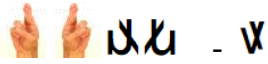
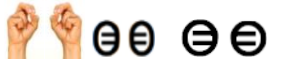
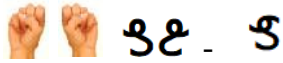
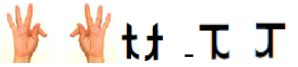
XAVIER, André Nogueira. **Descrição fonético-fonológica dos sinais da língua de sinais brasileira (Libras)**. 2006. 168 f. Dissertação (Mestrado) - Curso de Linguística, Universidade de São Paulo, São Paulo, 2006.

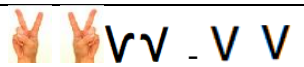
ANEXOS

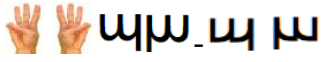
ANEXO A – Caracteres da SEL (2023)

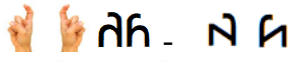
Formas digitais dos caracteres da escrita Sel
LETRAS:

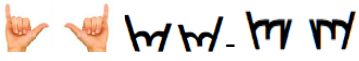
Quadro 1: Letras minúsculas e maiúsculas de configurações de mão

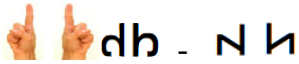
1. *a:* 	2. *legal:* 
3. *bê:* 	4. *bê-espraiado:* 
5. *mão espalmada:* 	6. *mão espraiada:* 
7. *cê:* 	8. *cê-encolhido:* 
9. *cê-espraiado:* 	10. *argola espraiada:* 
11. *gancho:* 	12. *concha:* 
13. *dê:* 	14. *dê-encolhido:* 
15. *e:* 	16. *garra encolhida:* 
17. *efe:* 	18. *pinça espraiada:* 
19. *gequê:* 	20. *pinça:* 
21. *agakapê:* 	22. *pinça dupla:* 
23. *ijota:* 	24. *ijota encolhido:* 
25. *ele:* 	26. *alicate:* 
27. *eme:* 	28. *garra:* 
29. *uene:* 	30. *uele:* 
31. *o:* 	32. *argola:* 
33. *grampo:* 	34. *pera:* 
35. *erre:* 	36. *argola dupla:* 
37. *esse:* 	
39. *tê:* 	
41. *namoro:* 	

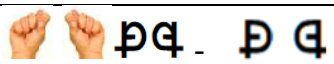
43. *vê:* 

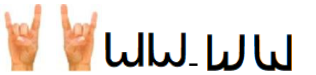
45. *dáblio:* 

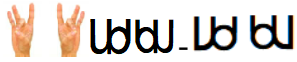
47. *xis:* 

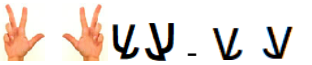
49. *ípsilon:* 

51. *zê:* 

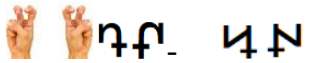
38. *figa:* 

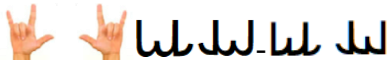
40. *chifre:* 

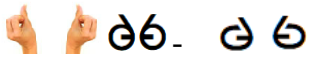
42. *argola média:* 

44. *vê-ele:* 

46. *desabrochar:* 

48. *cinco:* 

50. *love:* 

52. *seis:* 

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 2: Letras de partes do corpo

cabelo ,	dente  ,	tórax  ,
cabeça ,	língua  ,	barriga  ,
testa ,	queixo  ,	perna  ,
rosto ,	pescoço  ,	joelho  ,
sobrancelha ,	nuca  ,	axila  ,
olho  ,	ombro  ,	pálpebra  ,
nariz ,	costas  ,	lábio superior ,
bochecha ,	braço inteiro  ,	lábio inferior  ,
orelha  ,	braço  ,	nádegas 
buço ,	cotovelo ,	
boca ,	antebraço ,	
	punho ,	

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 3: Letras de dedos e formas combinadas de dedos

MÃO DIREITA
polegar (1), indicador (2), médio (3), anelar(4), mínimo (5).
poin (1-2), pomé (1-3), poan (1-4) , pomi (1-5) imé (2-3), mean (3-4), imi (2-5), imean (2-3-4), poimé (1-2-3), pomean (1-3-4), poami (1-4-5)

 poimean (1-2-3-4),  pomeami (1-3-4-5)  imeami (2-3-4-5),  poimeami (1-2-3-4-5).
MÃO ESQUERDA: polegar inv.  (1), indicador inv.  (2), médio inv.  (3), anelar inv.  (4), mínimo inv.  (5)
poin inv.  (2-1), pomé inv. (3-1), poan inv. (4-1) , pomi inv. (5-1) imé inv. (3-2), mean inv. (4-3), imi inv. (5-2), imean inv. (4-3-2), poimé inv. (3-2-1), pomean inv. (4-3-1), poami inv. (5-4-1) poimean inv. (4-3-2-1), pomeami inv. (5-4-3-1) imeami inv. (5-4-3-2), poimeami inv. (5-4-3-2-1).

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

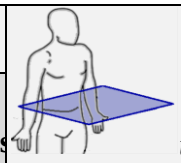
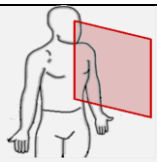
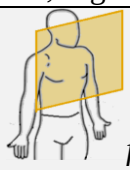
Quadro 4: Letras de movimentos retilíneos

Retilíneos
  para frente,  para trás,
   para cima,  para baixo,  para direita,  para esquerda,
  repetidas vezes para a direita,  repetidas vezes para esquerda,  direita e esquerda
 	
Retilíneos brevíssimos
  para frente,  para trás,
  para cima,  para baixo,  para direita,  para esquerda,
 repetidas vezes para a direita, repetidas vezes para esquerda , direita e esquerda.
 	
Retilíneos contidos
  para frente, para trás,
  para cima, para baixo, para direita, para esquerda,
 repetidas vezes para a direita, repetidas vezes para esquerda, direita e

esquerda

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 5: Letras de movimentos de mão nos plano transversal, sagital e frontal

Planos➔	 *transversal*				 *sagital*				 *frontal*			
Tipos de movimentos ➔												
	Direção ➔	indefi-nida	115ora-ria	anti-horá-ria		indefi-nida	115ora-ria	anti-horá-ria		indefi-nida	115ora-ria	anti-horá-ria
circular →												
	Direção ➔ ➔	p. frente		p. trás		p. frente		p. trás		p. cima		p. baixo
semicircular →	p.esq.				p.cima				p.esq.			
	p.dir.				p.baixo				p.dir.			
curvo →	p.esq.				p.cima				p.esq.			
	p.dir.				p.baixo				p.dir.			
angular →	p.esq.				p.cima				p.esq.			
	p.dir.				p.baixo				p.dir.			
angular duplo →	p.esq.				p.cima				p.esq.			
	p.dir.				p.baixo				p.dir.			
diagonal →	p.dir.				p.cima				p.dir.			
	p.esq.				p.baixo				p.esq.			
sinuoso →												
	p.dir.		p.esq.		p.cima		p.bai.		p.dir.		p.esq.	
zigue-zague →												
	p.dir.		p.esq.		p.cima		p.bai.		p.dir.		p.esq.	

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 6: Letras de movimentos de mão fora de plano

- *giro de punho:* para um lado e para o outro; em sentido horário; em sentido anti-horário;
- *dobradura de punho:* para o lado da palma; para o lado do dorso, para o lado do polegar; para o lado do dedo mínimo;
- *batida:* ;
- *tremura:* ;
- *inversão de palma:* .

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

DIACRÍTICOS:

Quadro 7: Diac.G1 – Eixo da mão e orientação de palma

<table>
<tr><td rowspan="6">Eixo da mão |EM|; orientação de palma |OP|</td><td colspan="4">Eixo superior:</td></tr>
<tr><td>para frente:
  </td><td>para trás:
  </td><td>para medial:
  </td><td>para lateral:
  </td></tr>
<tr><td colspan="4">Eixo medial/lateral:</td></tr>
<tr><td>para frente:
  </td><td>para trás:
  </td><td>para cima:
  </td><td>para baixo:
  </td></tr>
<tr><td colspan="4">Eixo anterior:</td></tr>
<tr><td>para cima:
  </td><td>para baixo:
  </td><td>para medial:
  </td><td>para lateral:
  </td></tr>
<tr><td>Eixos invertidos</td><td colspan="4">Eixo superior:
para frente: A A; para trás: ∧ ∧; para medial: ⊏ ⊐; para lateral: ⊐ ⊏.
Eixo medial/lateral:
para frente: ⋗ ⋖; para trás: ⋗ ⋖; para cima: ∪ ∪; para baixo: ∩ ∩.
Eixo anterior:
para cima: ⏚ ⏚; para baixo: ⍑ ⍑; para medial: ⏢ ⏢; para lateral: ⏢ ⏢.</td></tr>
</table>

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 8: Diac.G2 – Toque/proximidade à mão ou a partes do corpo

±	Palma da mão ou dedo ou lado da frente da parte do corpo;
± ⊥	dorso da mão ou dedo ou lado detrás da parte do corpo;
•	ponta(s) de dedo(s);
∝	lado do dedo polegar;
⊸	lado do dedo mínimo;
♈	entre dedos (ou, eventualmente, entre partes do corpo);
⭘	em volta de dedo(s);
⊚	em volta da mão ou de parte(s) do corpo;
∨	parte inferior da mão (punho) ou da parte do corpo;
<	à esquerda (de partes do corpo);
>	à direita (de partes do corpo);
∧	parte superior do corpo ou lado da mão correspondente aos dedos, mesmo com dedos encolhidos.

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 9: Diac.G3 – Posicionamento das duas mãos

- Para sinais com mãos na posição básica (mão esquerda do lado esquerdo e mão direita do lado direito) não se coloca nenhum diacrítico de posicionamento das mãos entre as letras de configuração de mão.
- Para sinais com mãos fora da posição básica, coloca-se entre as letras de configuração de mão, em:
  1- mãos alinhadas uma à frente da outra (qualquer mão) – diacrítico: (esquerda à frente: , direita à frente:);
  2- mãos alinhadas uma acima da outra (qualquer mão) – diacrítico: (esquerda acima: , direita acima:);
  3- mãos em diagonal no plano transversal (qualquer mão à frente da outra) – diacrítico:  (esquerda à frente:  , direita à frente: );
  4- mãos em diagonal no plano sagital (qualquer mão acima e à frente da outra) – diacrítico: (esquerda acima e à frente: , direita acima e à frente);
  5- mãos em diagonal no plano frontal (qualquer mão acima da outra) – diacrítico: (esquerda acima , direita acima);
  6- mãos cruzadas – diacrítico: .

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 10: Diac.G4 – Ordenamento de toque/proximidade em partes do corpo

Para sinais com partes do corpo tocadas por mais de uma mão, coloca-se entre letras de partes do corpo: 1 – partes do corpo tocadas ao mesmo tempo pelas duas mão – diacrítico ; 2 – partes do corpo tocadas alternadamente pelas duas mãos – diacrítico .

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 11: Diac.G5 – Expressão facial

Plásticas	Psicológicas
uma bochecha inflada; bochechas comprimidas; bochechas infladas; abrindo olhos; fechando olho ou olhos; dentadas / mordida / apertar entre dentes/ mastigar; soprando; sugando;	alegre/ feliz/ animado/ satisfeito/ esperançoso/ corajoso /rindo/ gargalhando; com medo/ assustado/ nervoso (ansioso); enojado/ insatisfeito/ com desprezo/ orgulhoso; irônico/ esperto/ malandro; triste/ deprimido/ com dor/ angustiado/ preocupado/ penalizado/ com dificuldade/ testa franzida; zangado/ nervoso (aborrecido)/ com ódio, testa franzida; surpreso/ boquiaberta/ abobalhado/ boca aberta/ bocejando.

gosto azedo/ fazendo bico; movimento de lábios ou língua/ oclusão com lábios; zigue-zague de queixo.	Gramaticais[24] sinal negativo; aspecto contínuo; grau aumentativo/ de intensificação; grau diminuitivo/ de suavização.

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 12: Diac.G6 – Tipos de movimentos de dedos

abrir gradativamente; abrir; abrir mais de uma vez; abrir lateralmente; abrir e fechar; dobrar dedo; dobrar dedo mais de uma vez;	fechar gradativamente; fechar; fechar mais de uma vez; fechar lateralmente; movimento tesoura; zigue-zague; esfregar.

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 13: Diac.G7 – Identificadores de movimentos retilíneos brevíssimos e contidos

Diacríticos colocados nas letras de movimento retilíneos, marcando: 1 – retilíneo brevíssimo – diacrítico: (movimento curtíssimo, quase zero); 2 – retilíneo contido – diacrítico: (movimento retilíneo que se detém marcadamente em algum ponto).

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 14: Diac.G8 – Composição dos movimentos das duas mãos

Para sinais com movimentos realizados com as duas mãos, coloca-se entre letras de movimento: 1 – movimento conjunto – diacrítico: (dois pontos entre as letras movimento); 2 – movimento alternado – diacrítico: (um ponto entre as letras movimento).

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

OUTROS CARACTERES:

[24] Em versões anteriores da Sel havia um diacrítico para ‘expressão facial interrogativa’ que foi eliminado para evitar redundância, pois verificarmos ser desnecessário, uma vez que essa expressão é também marcada pela pontuação. Quanto ao diacrítico de ‘aspecto contínuo’, que não se apresenta ainda claramente como expressão facial gramatical distintiva que necessite de uma marcação por diacrítico, esse permanece no quadro dos diacríticos de expressão facial gramatical, mesmo sem termos a certeza de sua necessidade, pois não há nenhum impedimento para seu desuso e eliminação futura, caso não se mostre necessário.

Quadro 15: Intensificação ou realização lenta do sinal

A intensificação ou realização lenta do sinal são indicadas por dois caracteres que são acrescentados ao final do sinal: ǁ Para desempenho intensificado/acelerado. ǀ Para desempenho mais lento.

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 16: Símbolos para numerais ordinais e para dinheiro

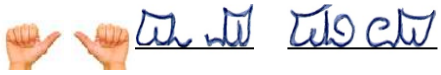
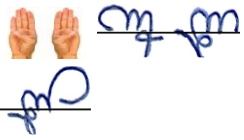
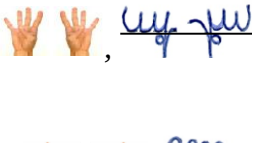
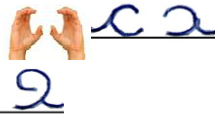
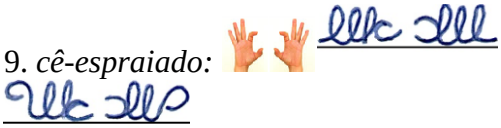
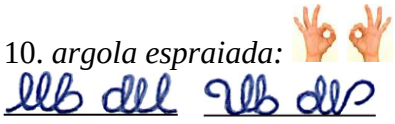
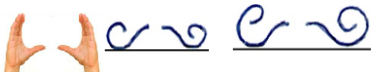
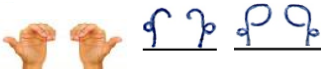
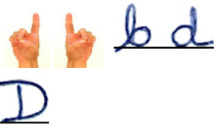
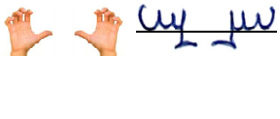
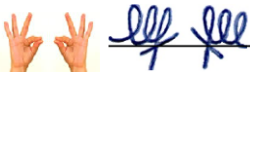
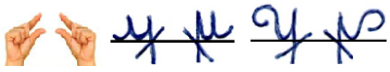
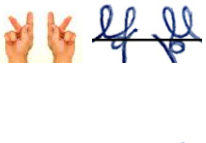
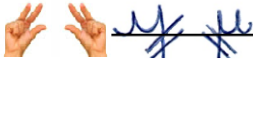
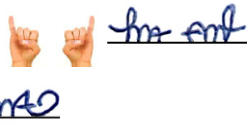
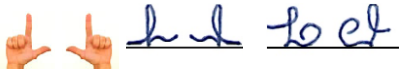
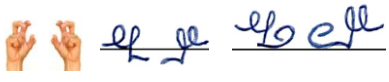
ª Para numerais ordinais (1ª, 2ª etc.). ₰ Para indicação de moeda (₰100,00 - Cem Reais). Escreve-se com a letra inicial do nome da moeda + ₰.

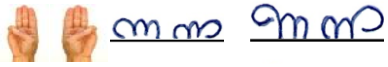
Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

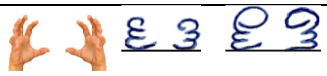
Formas manuscritas dos caracteres da escrita Sel

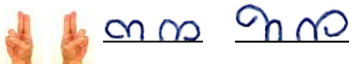
LETRAS:

Quadro 17: Letras manuscritas minúsculas e maiúsculas de configurações de mão (mãos esquerda e direita)

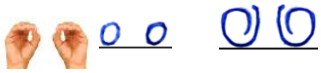
1. *a:* 	2. *legal:* 
3. *bê:* 	4. *bê-espraiado:* 
5. *mão espalmada:* 	6. *mão espraiada:* 
7. *cê:* 	8. *cê-encolhido:* 
9. *cê-espraiado:* 	10. *argola espraiada:* 
11. *gancho:* 	12. *concha:* 
13. *dê:* 	14. *dê-encolhido:* 
15. *e:* 	16. *garra encolhida:* 
17. *efe:* 	18. *pinça espraiada:* 
19. *gequê:* 	20. *pinça:* 
21. *agakapê:* 	22. *pinça dupla:* 
23. *ijota:* 	24. *ijota encolhido:* 
25. *ele:* 	26. *alicate:* 

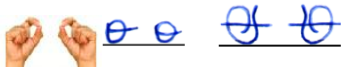
27. *eme:* 

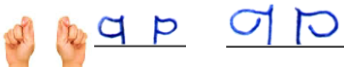
28. *garra:* 

29. *uene:* 

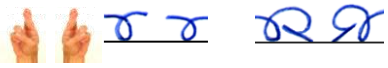
30. *uele:* 

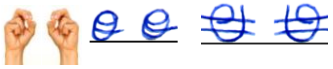
31. *o:* 

32. *argola:* 

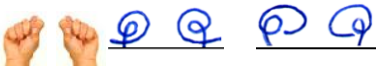
33. *grampo:* 

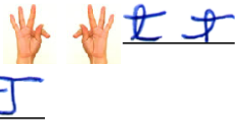
34. *pera:* 

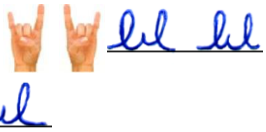
35. *erre:* 

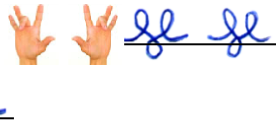
36. *argola dupla:* 

37. *esse:* 

38. *figa:* 

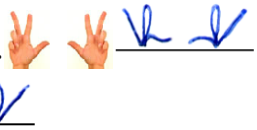
39. *tê:* 

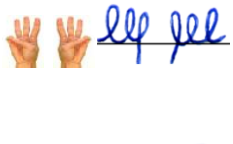
40. *chifre:* 

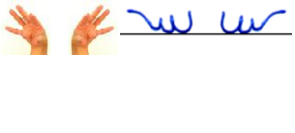
41. *namoro:* 

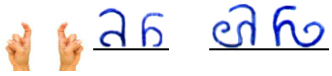
42. *argola média:* 

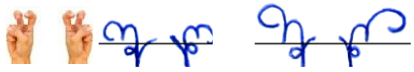
43. *vê:* 

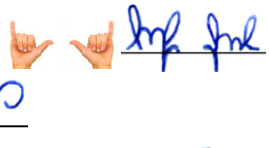
44. *vê-ele:* 

45. *dáblio:* 

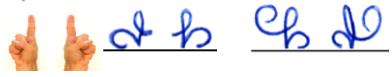
46. *desabrochar:* 

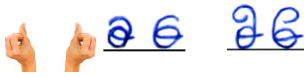
47. *xis:* 

48. *cinco:* 

49. *ípsilon:* 

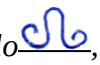
50. *love:* 

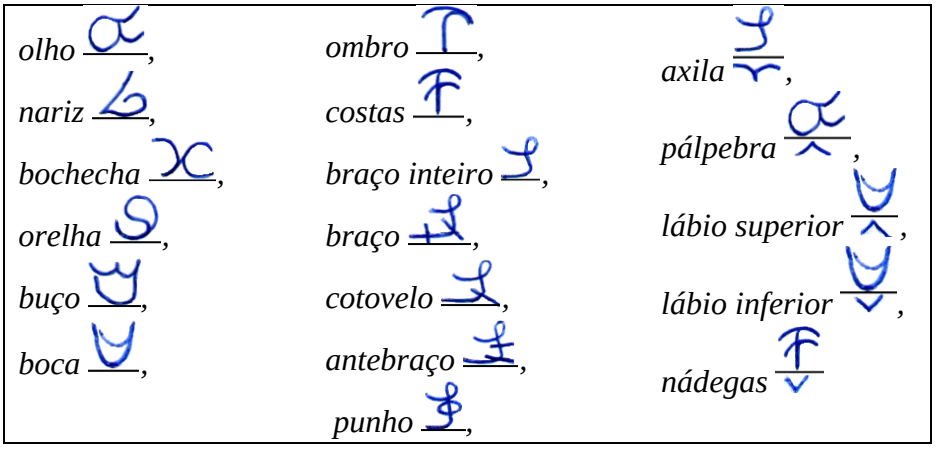
51. *zê:* 

52. *seis:* 

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 18: Letras manuscritas de partes do corpo

cabelo  ,	*dente*  ,	*tórax*  ,
cabeça  ,	*língua*  ,	*barriga*  ,
testa  ,	*queixo*  ,	*perna*  ,
rosto  ,	*pescoço*  ,	*joelho*  ,
sobrancelha  ,	*nuca*  ,	



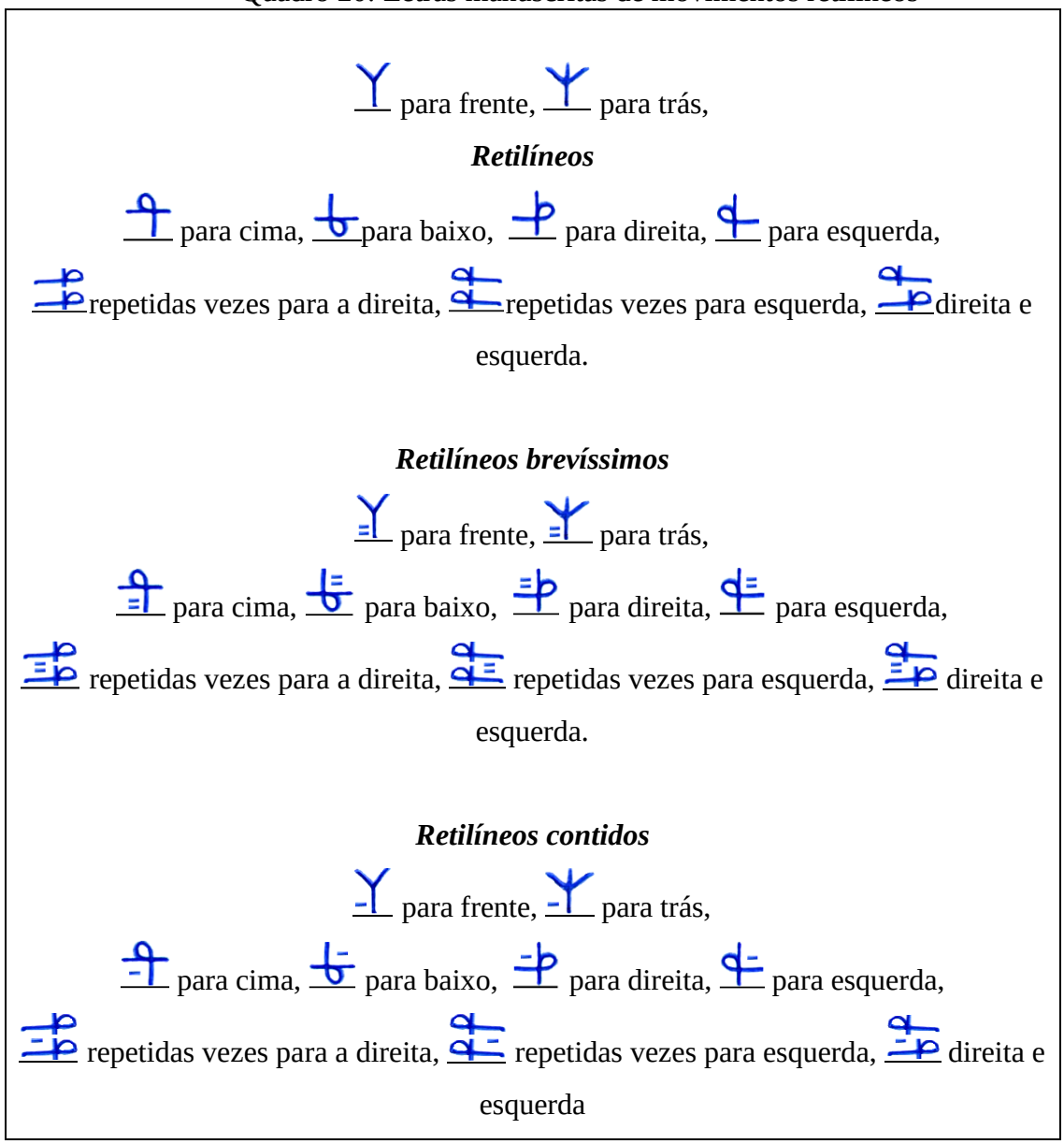
Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 19: Letras manuscritas de dedos e formas combinadas de dedos

MÃO DIREITA *polegar (1), indicador (2), médio (3), anelar(4), mínimo (5).*
poin (1-2), pomé (1-3), poan (1-4) , pomi (1-5) *imé (2-3), mean (3-4), imi (2-5),* *imean (2-3-4),* *poimé (1-2-3), pomean (1-3-4), poami (1-4-5)* *poimean (1-2-3-4), pomeami (1-3-4-5)* *imeami (2-3-4-5),* *poimeami (1-2-3-4-5).*
MÃO ESQUERDA: *polegar inv. (1), indicador inv. (2), médio inv. (3), anelar inv. (4),* *mínimo inv. (5).*
Poin inv. (2-1), pomé inv. (3-1), poan inv. (4-1) , pomi inv. (5-1) *imé inv. (3-2), mean inv. (4-3), imi inv. (5-2),* *imean inv. (4-3-2),* *poimé inv. (3-2-1), pomean inv. (4-3-1), poami inv. (5-4-1)* *poimean inv. (4-3-2-1), pomeami inv. (5-4-3-1)* *imeami inv. (5-4-3-2),* *poimeami inv. (5-4-3-2-1).*

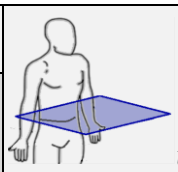
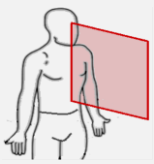
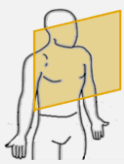
Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 20: Letras manuscritas de movimentos retilíneos



Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 21: Letras manuscritas de movimentos de mão nos planos transversal, sagital e frontal

Planos➔ / **Tipos de movimento** ↓		 *transversal*			 *sagital*			 *frontal*		
	Direção ➔	indefinida	horária	anti-horária		indefinida	horária	anti-horária		indefinida	horária	anti-horária

circular →												
	Direção ↓ →	p. frente		p. trás		p. frente		p. trás		p. cima		p. baixo
Semicircular →	p.esq.				p.cima				p.esq.			
	p.dir.				p.baixo				p.dir.			
curvo →	p.esq.				p.cima				p.esq.			
	p.dir.				p.baixo				p.dir.			
angular →	p.esq.				p.cima				p.esq.			
	p.dir.				p.baixo				p.dir.			
angular duplo →	p.esq.				p.cima				p.esq.			
	p.dir.				p.baixo				p.dir.			
diagonal →	p.dir.				p.cima				p.dir.			
	p.esq.				p.baixo				p.esq.			
sinuoso →												
	p.dir.		p.esq.		p.cima		p.bai.		p.dir.		p.esq.	
zigue-zague →												
	p.dir.		p.esq.		p.cima		p.bai.		p.dir.		p.esq.	

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 22: Letras de movimentos de mão fora de plano

- *giro de punho:*

para um lado e para outro; em sentido h

orário;

em sentido anti-horário;

- *dobrar dur*

a

de

punho:

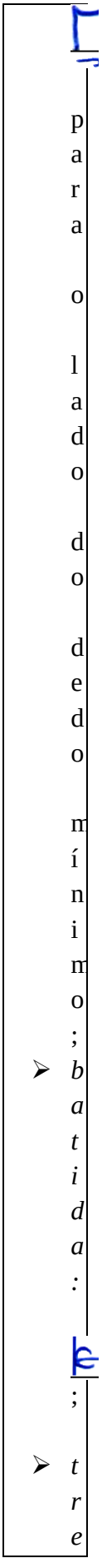
٢

para o lado da palma;

٢١

para o

lado do dorso, ☐ para o lado do pulegar;



- para o lado do dedo mínimo;
- *batida*;
- *tre*

mura:  ;

- *inversão de palma:*  .

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

DIACRÍTICOS:

Quadro 23: Diac.G1 – Eixo da mão e orientação de palma

Eixo superior:

para frente: , invertidos:  ; para trás: , invertidos: ;
para medial: , invertidos: ; para lateral: , invertidos:  .

Eixo medial:

para frente: , invertidos:  ; para trás: , invertidos:  ;
para cima: , invertidos: ; para baixo: , invertidos:  .

Eixo anterior:

para cima:  , invertidos:  ; para baixo:  , invertidos:  ;
para medial: , invertidos: ; para lateral: , invertidos: .

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 24: Diac.G2 – Toque/proximidade à mão ou a partes do corpo

± palma da mão ou dedo ou lado da frente da parte do corpo; ± ⊥ dorso da mão ou dedo ou lado detrás da parte do corpo; • ponta(s) de dedo(s); ∝ lado do dedo polegar; ʃ lado do dedo mínimo; ɣ entre dedos (ou, eventualmente, entre partes do corpo); ○ em volta de dedo(s); ⊕ em volta da mão ou de parte(s) do corpo; ∨ parte inferior da mão (punho) ou da parte do corpo; < à esquerda (de partes do corpo); > à direita (de partes do corpo); ^ parte superior do corpo ou lado da mão correspondente aos dedos, mesmo com dedos encolhidos; <> lados direito e esquerdo de partes do corpo.

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 25: Diac.G3 – Posicionamento das duas mãos

• Para sinais com mãos na posição básica (mão esquerda do lado esquerdo e mão direita do lado direito) não se coloca nenhum diacrítico de posicionamento das mãos entre as letras de configuração de mão. • Para sinais com mãos fora da posição básica, coloca-se entre as letras de configuração de mão, em: 6- mãos alinhadas uma à frente da outra (qualquer mão) – diacrítico: − (esquerda à frente: •−, direita à frente: −•); 2- mãos alinhadas uma acima da outra (qualquer mão) – diacrítico: | (esquerda acima: •|, direita acima: |•); 3- mãos em diagonal no plano transversal (qualquer mão à frente da outra) – diacrítico: ∠ (esquerda à frente: •∠, direita à frente: ∠•); 4- mãos em diagonal no plano sagital (qualquer mão acima e à frente da outra) – diacrítico: ⊬ (esquerda acima e à frente: •⊬, direita acima e à frente ⊬•); 5- mãos em diagonal no plano frontal (qualquer mão acima da outra) – diacrítico: ⁄ (esquerda acima •⁄, direita acima ⁄•); 6- mãos cruzadas – diacrítico: X.

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 26: Diac.G4 – Ordenamento de toque/proximidade em partes do corpo

Para sinais com partes do corpo tocadas por mais de uma mão, coloca-se entre letras de partes do corpo: 1 – partes do corpo tocadas ao mesmo tempo pelas duas mão – diacrítico ••;

2 – partes do corpo tocadas alternadamente pelas duas mãos – diacrítico .

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 27: Diac.G5 – Expressão facial

Plásticas	Psicológicas
uma bochecha inflada; bochechas comprimidas; bochechas infladas; abrindo olhos; fechando olhos; dentadas / mordida / apertar entre dentes/ mastigar; soprando; sugando; gosto azedo/ fazendo bico; movimento de lábios ou língua/ oclusão com lábios; zigue-zague de queixo.	alegre/ feliz/ animado/ satisfeito/ esperançoso/ corajoso/ rindo/ gargalhando; com medo/ assustado/ nervoso (ansioso); enojado/ insatisfeito/ com desprezo/ orgulhoso; irônico/ esperto/ malandro; triste/ deprimido/ com dor/ angustiado/ preocupado/ penalizado/ com dificuldade/ testa franzida; zangado/ nervoso (aborrecido)/ com ódio, testa franzida; surpreso/ boquiaberta/ abobalhado/ boca aberta/ bocejando.
	Gramaticais
	sinal negativo; aspecto contínuo; grau aumentativo/ de intensificação; grau diminutivo/ de suavização;

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 28: Diac.G6 – Tipos de movimentos de dedos

abrir gradativamente; abrir; abrir mais de uma vez; abrir lateralmente; abrir e fechar; dobrar dedo; dobrar dedo mais de uma vez;	fechar gradativamente; fechar; fechar mais de uma vez; fechar lateralmente; movimento tesoura; zigue-zague; esfregar.

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 29: Diac.G7 – Identificadores de movimentos retilíneos brevíssimos e contidos

Diacríticos colocados nas letras de movimento retilíneos, marcando: ➢ retilíneo brevíssimo: (movimento curtíssimo, quase zero); ➢ retilíneo contido: (movimento retilíneo que se detém marcadamente em algum ponto).

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

Quadro 30: Diac.G8 – Composição dos movimentos das duas mãos

Para sinais com movimentos realizados com as duas mãos, coloca-se entre letras de movimento: ➢ movimento conjunto: ⠒ (dois pontos entre as letras movimento); ➢ movimento alternado: ⠂ (um ponto entre as letras movimento).

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

OUTROS CARACTERES:

Quadro 31: Intensificação ou realização lenta do sinal

A intensificação ou realização lenta do sinal são indicadas pelos seguintes caracteres, que são acrescentados ao final do sinal: ⠇⠇⠇ Para desempenho intensificado/acelerado. ⠇⠇ Para desempenho mais lento.

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

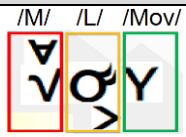
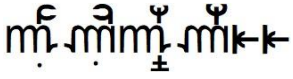
Quadro 32: Símbolos para numerais ordinais e para dinheiro

⠪ Para numerais ordinais ⠈⠑Para indicação de moeda

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

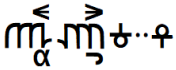
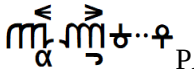
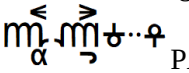
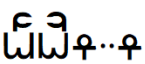
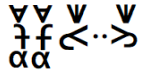
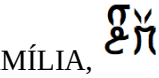
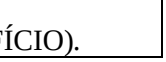
ANEXO B – Regras da SEL (2023)

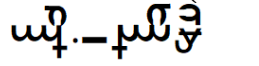
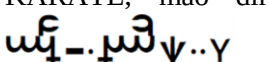
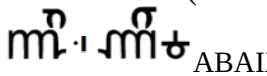
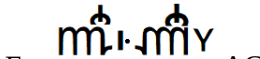
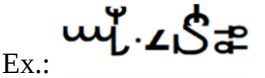
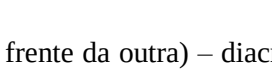
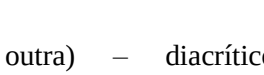
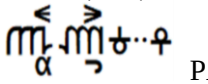
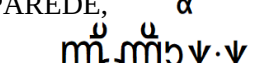
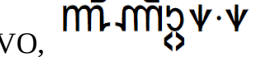
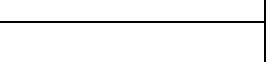
Regras da escrita Sel: referentes à unidade MLMov

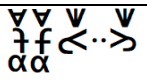
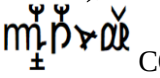
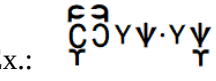
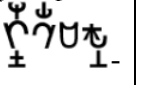
Descrição da regra	REGRA
1. Formação da unidade MLMov	As letras e diacríticos representam os macrossegmentos, ordenados obrigatoriamente da esquerda para a direita como: Mão /M/, Locação /L/ e Movimento /Mov/.  VER
2. Sinais com mais de uma unidade MLMov	Sinais com mais de uma unidade MLMov são escritos preservando-se a ordem de realização dessas unidades sem deixar nenhum espaço em branco entre elas ( ESCOLA).

Fonte: Quadro produzido com base nas regras da Sel apresentada em Lessa-de-Oliveira (2023)

Regras da escrita Sel: referentes ao macrossegmento Mão /M/

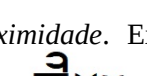
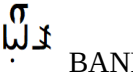
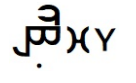
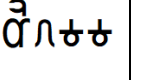
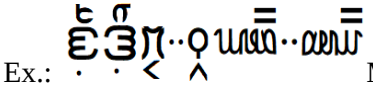
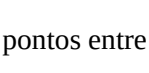
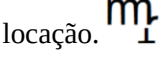
Descrição da regra	REGRA
3. Representação da CM	No macrossegmento /M/, apenas o traço *configuração de mão* |CM| é representado por uma letra, na forma minúscula ou maiúscula. As letras maiúsculas ocorrem em nomes próprios e em início de frases. Em sinais realizados com as duas mãos com configurações idênticas, as letras CM de ambas as mãos podem opcionalmente ficar na forma maiúscula.  PAREDE ou  PAREDE.
4. Diacrítico Eixo-Palma	Os traços de *eixo da mão* |EM| e *orientação de palma* |OP| são representados pelo mesmo diacrítico, obrigatoriamente colocado sobre a letra de |CM|.  PAREDE,  EDIFÍCIO.
5. Diacrítico Eixo-Palma no movimento	Havendo, durante o movimento, uma mudança de *eixo da mão* |EM| e/ou *orientação de palma* |OP| que cause problema de processamento, se não indicada, os diacríticos *Eixo-palma* são também colocados sobrescritos às letras de movimento, para marcar essa alteração. Ex.:  FAMÍLIA,  ÓDIO. *Obs.: Mais raramente, esse diacrítico pode também ocorrer sobre a letra da |PC| também marcando alteração do eixo da mão.* Ex.: SURDO
6. Diacrítico de Posição das duas mãos	O traço de *posicionamento das duas mãos* |PDM| é representado por um diacrítico de *posição das duas mãos*, colocado entre as letras de configuração das mãos direita e esquerda, da seguinte forma, em: • <u>mãos alinhadas em posição de base</u> (direita do lado direito e esquerda do lado esquerdo) – <u>NÃO</u> ocorre diacrítico de *posicionamento das duas mãos* ( EDIFÍCIO).

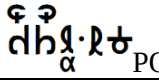
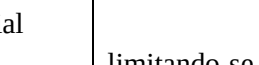
	• <u>mãos alinhadas uma à frente da outra</u> (qualquer mão) – diacrítico: ▬ , Ex.:  PROVA (mão esquerda à frente: diacrítico •▬, Ex.:  - KARATÊ; mão direita à frente: diacrítico ▬•, Ex.:  FILA); • <u>mãos alinhadas uma acima da outra</u> (qualquer mão) – diacrítico: ▮ (mão esquerda acima: diacrítico •▮, Ex.:  ABAIXO; mão direita acima: diacrítico ▮•, Ex.:  ACIMA); • <u>mãos em diagonal no plano transversal</u> (qualquer mão à frente da outra) – diacrítico: ◢, Ex.: GOLFINHO (mão esquerda à frente, diacrítico: •◢, Ex.:  VIOLINO; mão direita à frente, diacrítico: ◢•); • <u>mãos em diagonal no plano sagital</u> (qualquer mão acima e à frente da outra) – diacrítico: ◤ , Ex.:  CAÇAR (mão esquerda acima e à frente, diacrítico: •◤; mão direita acima e à frente, diacrítico: ◤•); • <u>mãos em diagonal no plano frontal</u> (qualquer mão acima da outra) – diacrítico: ◢, Ex.:  DESCONFIAR (mão esquerda acima, diacrítico: •◢; mão direita acima, diacrítico: ◢•); • <u>mãos cruzadas</u> – diacrítico: ✖, Ex.:  EXPLOSÃO . *Obs.: Como regra geral, não se usa diacrítico de posição das duas mãos, quando ocorre o diacrítico de toque/proximidade (a seguir), pois o uso desses dois tipos de diacríticos é redundante. Mas a coocorrência desses diacríticos pode acontecer em raríssimos casos, quando ambos se fazem necessários à identificação do sinal.*
7. Diacrítico toque/proximidade	Se necessário, diacríticos de *toque/proximidade* são colocados sob as letras de: *configuração de mão* |CM|, de *dedos* |D/CD| ou de *partes do corpo* |PC|. Ex.:  PARADE,  POR QUE/ PORQUE,  VIDA/ VIVO,  SERVIÇO/EMPREGADO DOMÉSTICO.
8. Diacrítico toque/ proximidade em sinais com as duas mãos	O diacrítico *toque/proximidade* é colocado: • <u>sob as |CM| das duas mãos</u>, se ambas se tocam antes do

	movimento - Ex.:  FAMÍLIA • <u>apenas sob a letra |CM| da mão de base</u> (esquerda para os destros), se o toque ocorre depois do movimento - Ex.:  COMPRAR • ou <u>sob a letra de movimento da mão direita</u>, se for muito necessário indicar onde se deu o toque na mão principal após o movimento - Ex.:  COMUNICAÇÃO,  - ABACATE

Fonte: Quadro produzido com base nas regras da Sel apresentada em Lessa-de-Oliveira (2023)

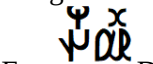
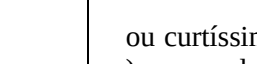
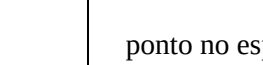
Regras da escrita Sel: referentes ao macrossegmento Locação /L/

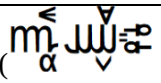
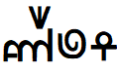
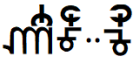
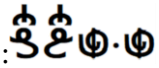
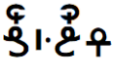
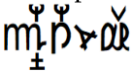
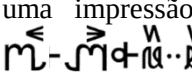
Descrição da regra	REGRA
9. Ausência do diacrítico toque/ proximidade na locação	Em *posições básicas* de locações (região central da frente das partes do corpo em geral, dorso do braço e do antebraço, lado da bochecha ou do cabelo correspondente ao da mão de principal), não se escreve o diacrítico de *toque/proximidade*. Ex.:  PESSOA,  BANHEIRO/SANITÁRIO,  MULHER,  PENTE.
10. Representação da PC	No macrossegmento /L/ apenas o traço parte do corpo |PC| é representado por uma letra, a qual não ocorre quando o sinal é realizado sem o envolvimento de uma parte do corpo. Ex.:  IDOSO ,  TRABALHO.
11. Diacrítico conjuntamente /alternado na locação	Em sinais com mais de uma |PC| tocadas pelas duas mãos, colocam-se: • *<u>dois pontos</u>*, se as |PC| forem <u>tocadas ao mesmo tempo</u> pelas duas mãos – Ex.:  MACACO • *<u>um ponto</u>*, se as |PC| forem <u>tocadas alternadamente</u> pelas duas . *Obs.: A |PC| representada do lado esquerdo desse diacrítico corresponde à parte do corpo tocada pela mão de base (esquerda para os destros) e a representada do lado direito corresponde à parte do corpo tocada pela mão principal (direita para os destros).*
12. Mais de uma PC tocada por uma mão	Em sinais com mais de uma |PC|, tocadas pela mesma mão, escreve-se as |PC| na ordem de toque, sem pontos entre elas -  SURDO
13. PC tocada necessariamente pela mão de base	Para sinais com |PC| tocada necessariamente pela mão de base, é obrigatória a ocorrência do diacrítico *um ponto* grafado após a letra de locação.  MADEIRA *Obs.: Quando a |PC| é tocada apenas pela mão principal esse diacrítico não ocorre.*

14. Letras de dedos em /L/	As letras que representam os dedos podem funcionar como elementos de /L/, se não houver movimento de dedos e eles estiverem servindo apenas para ancorar a realização do sinal.  PORQUE
15. Diacrítico Expressão Facial	Os diacríticos de *expressão facial* |ExpF| têm um uso muito restrito, limitando-se a sinais psicológicos ( CIÚME), gramaticais (NÃO PODE) ou com expressão facial plástica (CEGO).
16. Posição do diacrítico Expressão Facial	Os diacríticos de *expressão facial* ocorrem sobrescritos a: • uma letra de /L/ ( TRISTE) ou de /Mov/ (NÃO), caso não haja /L/, ou • encostado na parte superior da última letra do sinal ( GORDO), se /Mov/ não puder receber esse diacrítico ou não existir no sinal.

Fonte: Quadro produzido com base nas regras da Sel apresentada em Lessa-de-Oliveira (2023)

Regras da escrita Sel: referentes ao macrossegmento Movimento /Mov/

Descrição da regra	REGRA
17. Representação dos cinco dedos	Movimentos realizados por dedos isolados são representados pelas cinco letras que identificam os dedos *polegar, indicador, médio, anelar e mínimo*. ( polegar, indicador, médio, anelar, mínimo)
18. Combinação de dedos	Os movimentos que envolvem dedos em conjunto são representados por uma letra formada por combinações das letras dos cinco dedos. ( , , etc.)
19. Diacrítico Movimento de dedo	Os diacríticos de *tipos de movimento de dedo* ocorrem obrigatoriamente sobrescritos às letras de dedos. Ex.:  DINHEIRO, ANDAR/CAMINHAR
20. Representação de movimentos de mão retilíneos	Os movimentos de mão *retilíneos* são representados pelas letras que indicam sempre *tipo de movimento de mão* |TMovM| e *direção do movimento* |DMov|.  para frente, para trás, para cima, para baixo, , para direita, para esquerda, repetidas vezes para a direita, repetidas vezes para esquerda, direita e esquerda
21. Diacrítico Contido/brevíssimo de movimentos de mão retilíneos	Os movimentos retilíneo brevíssimo (com quase zero deslocamento) e retilíneo contido são representados obrigatoriamente pelo acréscimo dos diacríticos: - e = . O retilíneo contido corresponde a um movimento (longo, curto ou curtíssimo) que se detém em algum ponto ( CRIANÇA) e, quando combinado com outros movimentos, indica a direção da repetição do movimento que, nesse caso, não se dá no mesmo ponto no espaço.  (ANDAR de um prédio) O retilíneo brevíssimo não é um movimento curto, é um

	movimento com quase zero de deslocamento ( CINEMA)
22. Representação de movimentos de mão em planos	Os oito tipos de movimentos de mão do grupo *em planos* são representados sempre por letras que correspondem a três traços: • *tipo de movimento de mão* |TMovM| (definido pelo formato da cauda da letra:  diagonal,  zigue-zague,  sinuoso,  angular,  angular duplo,  curvo,  semicircular, ou da letra inteira:  circular); • *direção do movimento de mão* |DMov| (definida pelo formato ou posição da cabeça da letra:  p. frente,  p. trás,  p. esquerda,  p. direita, ou por diacrítico:  horário); e • *plano de movimento* |PMov| (definido por marca específica na cabeça da letra:  frontal,  sagital,  transversal, ou na letra inteira:  frontal,  sagital,  transversal.
23. Representação de movimentos de mão fora de planos	Os cinco movimentos de mão *fora de plano* são representados por sua letra específica. ( giro de punho,  dobradura de punho,  batida,  tremura,  inversão de palma).
24. Dígrafos	Alguns movimentos podem ser escritos com a combinação de duas letras, embora seja um só movimento (dígrafos). Ex.:  IMPORTANTE
25. Diacrítico de movimento conjunto	Para marcar *movimento conjunto* das duas mãos, coloca-se dois pontos entre as letras de movimento das mãos esquerda e direita. Ex.:  MESA
26. Diacrítico de movimento alternado	Para marcar *movimento alternado* entre as duas mãos, coloca-se apenas um ponto entre as letras de movimento das mãos esquerda e direita. Ex.:  BICICLETA
27. Movimento realizado por apenas uma das mãos	Para sinais com *movimento realizado apenas por uma das mãos*, escreve-se a letra de movimento sem ponto algum, independentemente de qual mão tenha realizado o movimento. Ex.:  GUARDA-CHUVA
28. Mais de um movimento realizado por uma das mãos	Para marcar *mais de um movimento realizado apenas por uma das mãos*, escreve-se as letras na ordem de realização dos movimentos (da direita para a esquerda), sem dividi-las com pontos. Ex.:  COMPRAR
29. Simetria nos movimentos das duas mãos	Havendo mais de um movimento realizado pelas duas mãos, esses são escritos com os caracteres de movimento da mão esquerda, e a ordem desses, invertidos em relação aos da mão direita (passando uma impressão visual com foco do centro para fora). Ex.:  VÁRIOS

Fonte: Quadro produzido com base nas regras da Sel apresentada em Lessa-de-Oliveira (2023)

Regras da escrita Sel: referentes a propriedade sintáticas e numéricas

Descrição da regra	REGRA
30. Intensificação ou realização lenta do sinal	A intensificação ou realização lenta do sinal são indicadas por dois caracteres que são acrescentados ao final do sinal: • ||| para desempenho intensificado/acelerado -  PEDALAR RÁPIDO • || para desempenho mais lento -  PEDELAR LENTAMENTE
31. Pontuação	A pontuação em Sel é basicamente a mesma utilizada para o Português, com exceção dos pontos de interrogação e exclamação, que são, como no espanhol, utilizados no início (¿) e no final da pergunta (?)e no início (¡) e final (!) da frase exclamativa.
32. Representação de moeda	A indicação de moeda é escrita com a letra inicial do nome da moeda + o símbolo  . Exemplo: - Cem Reais.
33. Representação de ordenais	A indicação de numerais ordinais realiza-se com o símbolo  , escrito ao lado dos algarismos. Exemplo - primeiro, - segundo etc.

Fonte: Quadro produzido com base nas regras da Sel apresentada em Lessa-de-Oliveira (2023)

Regra para escrita da datilologia utilizando caracteres da escrita Sel

Regra A

A representação da datilologia é feita apenas com as letras de *configuração da mão direita* escrita na mesma ordem da palavra soletrada, sem a utilização dos diacríticos de *eixo* e *orientação de palma*, utilizando-se os diacríticos  ou , para diferenciar as letras que são representadas pela mesma configuração de mão:

a: ; bê:  ; cê: ; cê-cedilha: ; dê: ; e: ; efe: ; gê: ; agá: ; i: ; jota ; cá: ; ele: ; eme: ; ene: ; o: ; pê ; quê ; erre: ; esse: ; tê: ; u: ; vê: ; dáblio: ; xis: ; ípsilon: ; zê: .

Regra B

Para representar os acentos do Português, utilizam-se os diacríticos de expressão facial a seguir:

Agudo: ; crase: ; circunflexo: ; til: ; trema: .

Fonte: Quadro produzido com base em Lessa-de-Oliveira (2023)

ANEXO C – Caracteres da ELIS

CONFIGURAÇÃO DE DEDOS	
Polegar	**Demais dedos**
. fechado	. fechado
⟋ na palma	⌝ muito curvo
⌜ curvo	⌝ curvo
⟍ "3D"	⟍ inclinado
_ horizontal	I estendido
I vertical	

Fonte: Barros (2016, p. 205-207)

ORIENTAÇÃO DA PALMA
palma para frente
palma para trás
palma para cima
palma para baixo
palma para a medial
palma para a distal

Fonte: Barros (2016, p. 205-207)

PONTO DE ARTICULAÇÃO			
Cabeça	**Tronco**	**Membros**	**Mão**
rosto	pescoço	braço inteiro	palma
alto da cabeça	corpo	ombro	dorso
lateral da cabeça	tórax	axila	dedos
orelha	ao lado do corpo	braço	lateral de dedo
testa	abdômen	cotovelo	intervalo de dedo
sobrancelha		antebraço	articulações
olho		punho	ponta de dedo
maçã do rosto		perna	
nariz			
buço			
boca			
dentes			
bochecha			
queixo			
abaixo do queixo			

Fonte: Barros (2016, p. 205-207)

MOVIMENTO					
Braço e punho		**Mão**		**Expressões não-manuais**	
⊥	para frente	⫫	abrir	⍝	negação com a cabeça
⊤	para trás	⫪	fechar	⍜	afirmação com a cabeça
⫩	para frente e para trás	⫬	abrir e fechar	⍥	língua na bochecha
↑	para cima	⫬	flex. dedos na base	⍦	língua para fora
↓	para baixo	⫬	flex. dedos na ponta	⍧	corrente de ar
↕	para cima e para baixo	⩛	unir e separar dedos	⍨	vibração dos lábios
→	para a direita	⩘	tamborilar de dedos	⍩	mov. lateral do queixo
←	para a esquerda	⩙	friccionar de dedos	⍪	murchar bochechas
↔	para a dir e a esq	⌙	dobrar o punho	⍫	inflar bochechas
+	para o meio	⌙	mov lateral do punho	⍬	abrir a boca
↔	para fora	⌙	girar o punho	÷	piscar
↗	para cima e à direita	⌙	girar antebraço	⍭	girar o tronco
↖	para cima e à esquerda				
↘	para baixo e à direita				
↙	para baixo e à esquerda				
∩	arco				
⊾	flex/ext de braço				
0	circular vertical				
o	circular horizontal				
⊙	circular frontal				

Fonte: Barros (2016, p. 205-207)