A transcrição de textos do Corpus de Libras

Ronice Müller de Quadros
Universidade Federal de Santa Catarina

Resumo

Este artigo apresenta as decisões tomadas para o estabelecimento dos procedimentos para a realização das transcrições básicas do Corpus de Libras e das anotações sugeridas para o desenvolvimento de análises mais específicas dos dados em Libras. Primeiramente, contextualizamos o Corpus de Libras e os projetos que o compreendem. A partir disso, apresentamos alguns estudos sobre transcrições de línguas de sinais que apresentam problemas e sugestões de encaminhamentos para resolvê-los, visto que são caracterizados a partir do caráter multimodal dos dados implicados em corpora de línguas de sinais. Considerando os estudos já realizados, apresentamos as decisões tomadas para o estabelecimento dos procedimentos de transcrição e anotação dos dados do Corpus de Libras, que resultaram em um Manual de Transcrição da Libras disponível na página do Corpus de Libras (www.corpuslibras.ufsc.br).

Palavras chaves: Corpus, Libras, Documentação, Anotação, Transcrição

Abstract

This paper describes the decisions made to establish procedures for carrying out basic transcriptions of the Libras Corpus and the annotations suggested to develop more specific analyzes of the Libras data. Firstly, we describe the Libras Corpus and the research projects conducted within it. We then describe some studies on transcribing sign languages that have presented problems caused by the multimodal character of the data in the sign languages corpora and offer some suggestions of how to solve the problems. The decisions taken to establish the procedures for transcription and annotation of the data in the Corpus have resulted in a Manual of Transcription of Libras available on the Libras Corpus page.

Key words: Corpus, Libras, Documentation, Annotation, Transcriptions

Contextualização

O que é corpora de línguas

Corpora de línguas são registros de diferentes gêneros textuais escritos e falados de diferentes línguas. Segundo MacCarthy e O'Keeffe (2010), a ideia de corpus de uma língua começou em 1960, com os estudos lexicográficos. Com o avanço da tecnologia, os corpora começaram a ser compilados com o uso de computadores. A exemplo disso, os autores citam o primeiro corpus do inglês, que foi coletado e organizado na Brown University, com um milhão de palavras do inglês de textos literários. A partir de 1970, já havia vários corpora de línguas sendo estabelecidos. MacCarthy e O'Keeffe (2010), mencionam que os corpora servem para disponibilizar dados linguísticos em quantidade para possibilitar a verificação de padrões que são usados. Assim sendo, a linguística de corpus viabiliza o acesso a grandes quantidades de dados para que os linguistas

possam explicar os fenômenos linguísticos. A tarefa do linguista passa, então, a ser a elaboração de metodologias confiáveis para descrever e dar conta dessas evidências linguísticas. Os autores ainda discutem a questão do tamanho do corpus. Há um movimento na linguística de corpus de megacorpus para minicorpus, no sentido de organização de um corpus específico para atender a determinados objetivos, conforme a proposta da pesquisa. Além disso, com a revolução da internet, mais e mais temos acesso a diferentes corpora, que vão de monomodal a multimodal, ou seja, os dados linguísticos incluem, além da informação verbal ou escrita, informações corporais, imagéticas, sonoras e, em algumas circunstâncias, até táteis e olfativas.

Os autores definem corpus como compêndio linguístico de um texto (*parole*) com evidências para a compreensão sobre a língua (*langue*) com dois objetivos centrais: (1) verificar a extensão de um padrão encontrado (valor descritivo) e (2) analisar os fatores contextuais que influenciam a variabilidade (valor explicativo). Esses objetivos exigem a identificação e a análise de ocorrências no uso da língua e, para conclusões mais abrangentes sobre um fenômeno linguístico, é necessária uma grande quantidade de dados de diferentes usuários da língua. Esses são alguns dos problemas metodológicos levantados por MacCarthy e O'Keeffe.

No presente artigo, estaremos considerando corpus como um banco de produções de uma língua, organizado por diferentes tipos de textos (orais, sinalizados e/ou escritos), para fins de registro da língua. No caso específico do corpus em questão, o Corpus de Libras, temos vários conjuntos de produções da Libras organizados a partir de projetos de pesquisa com diferentes propostas, mas todos têm em comum o registro de interações em Libras por meio de filmagens em vídeo. Alguns deles possuem também transcrições e traduções.

Histórico do Corpus de Libras

O Corpus da Língua Brasileira de Sinais (Libras)[1] começou a ser constituído em 1995. Este Corpus envolve diferentes projetos, compreendendo dados de fontes diversas e diretrizes para o registro dos dados e metadados em Libras. O primeiro deles documenta dados de estudos longitudinais com crianças surdas filhas de pais surdos adquirindo a Libras. Posteriormente, foram incluídas crianças surdas filhas de pais ouvintes; crianças surdas com implante coclear, filhas de pais surdos e pais ouvintes; e crianças ouvintes filhas de pais surdos (Codas).

Atualmente temos os seguintes dados organizados no banco de dados relativo ao desenvolvimento da Libras:

Tabela 1: Corpus de Libras de Aquisição da Linguagem

Dados Projeto de Pesquisa						
Sujeito	**Período de coleta**	**Vídeos de cada sessão**	**Transcrições realizadas**	**Filmagem**	**Contexto**	**Financiamento**
Ana	01;01 a 03;03	30 sessões	30 sessões transcritas	Concluída	Surda/ pais surdos	CAPES
Leo	01;07 a 03;11	78 sessões	21 sessões transcritas	Concluída	Surdo/ pais surdos	CAPES, CNPQ
Bia	03;05 a 06;05	103 sessões	17 sessões transcritas	Concluída	Surda/ pais ouvintes	CNPQ
Igor	02;01 a 03;05	64 sessões	30 sessões transcritas	Concluída	Coda	CNPQ, NIH
Edu	00;09 a 04;00	62 sessões	35 sessões transcritas	Concluída	Coda	CNPQ, NIH
Bruno	01;02 a 04;09	93 sessões	14 sessões transcritas	Concluída	IC/pais surdos	CNPQ, NIH
Tainá	02;01 a 04;07	55 sessões	11 sessões transcritas	Concluída	IC/pais ouvintes	CNPQ

[1] O Corpus de Libras está sendo constantemente alimentado e encontra-se disponível no Portal de Libras, www.libras.ufsc.br, no link do corpus que pode ser acessado diretamente em www.corpuslibras.ufsc.br. Os dados disponibilizados envolvem projetos que contaram com diferentes fontes de fomento: CNPQ, CAPES, IPHAN e NIH.

O Banco de Dados de aquisição da Libras atingiu maturidade metodológica, pois desenvolveu uma série de ferramentas que possibilitaram a organização dos dados para a realização das análises. Os dados estão organizados de tal forma a viabilizar pesquisas por terceiros, contando sempre com as devidas precauções observadas pelo Comitê de Ética. O acesso aos dados começa a ser mais amplo, mas mantém-se restrito no sentido de resguardar a visualização dos vídeos de crianças. Para isso, o Comitê de Ética exigiu cartas de consentimento que especificassem a permissão explícita dos pais e, quando possível, da própria criança, para o acesso irrestrito a demais pesquisadores. O acesso irrestrito enriquece o próprio banco de dados e as produções de pesquisa, que se multiplicam, uma vez que permite a inclusão de transcrições adicionais, bem como de análises dos dados que constituem as pesquisas no próprio banco de dados, consolidando o Corpus de Libras de aquisição da linguagem.

A metodologia estabelecida no escopo desse banco de dados serviu de referência para os materiais que estão sendo usados no Inventário Nacional de Libras que integra o Corpus de Libras. Esse inventário tem o objetivo de estabelecer a documentação da Libras em âmbito nacional e já conta com dados coletados da Grande Florianópolis, (Santa Catarina) e, em fase de coleta, de Maceió (Alagoas)[2]. Essa coleta de dados objetiva ser replicada em todo o Brasil para o estabelecimento de um Corpus da Libras com dados que permita análises comparáveis da Libras de diferentes regiões do país. A metodologia usada para o Inventário Nacional de Libras compreende interações de surdos em pares divididos em três grupos, por idade e por gênero. Todos os procedimentos para a coleta dos dados, organização dos dados e metadados e transcrição dos dados foram aplicados e ajustados para serem usados em todo o país e permitirem dados em Libras comparáveis entre os surdos de diferentes regiões.

[2] O Inventário de Libras da Grande Florianópolis está sob a coordenação de Ronice Müller de Quadros e conta com o financiamento do CNPQ (Processos 303725/2013-3 e 471355/2013-5), e o Inventário de Libras de Maceió está sob a coordenação de Jair Barbosa da Silva, com o financiamento do CNPQ (460589/2014-8).

Assim, com os dados seguindo os mesmos procedimentos metodológicos, teremos condições de analisá-los para identificar os fatores contextuais que influenciam a variabilidade da Libras, explicando os fenômenos linguísticos estudados.

O Corpus de Libras também inclui dados do Libras Acadêmico, que compreende produções do Exame ProLibras, exame de proficiência e certificação de Libras do Ministério da Educação e do Curso de Letras Libras EAD, ambos implementados pela Universidade Federal de Santa Catarina. O Libras Acadêmico inclui a publicação de vários materiais produzidos pelos alunos durante o oferecimento do curso, especialmente trabalhos realizados em Libras e literatura em Libras. Os materiais disponibilizados contam com a permissão direta de ex-alunos e dos participantes das atividades do Exame ProLibras. Os materiais coletados foram catalogados e publicados no Corpus de Libras Acadêmico no respectivo estado do polo, no qual o aluno estava atendendo ao curso. Esses materiais são muito úteis para alunos surdos, alunos interessados em Libras, professores de Libras e educadores bilíngues. Essa documentação está sendo complementada por meio do Inventário Nacional de Libras, com o financiamento do Ministério da Cultura, pelo Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN), por meio de uma parceria com o Instituto de Políticas Linguísticas (IPOL) e a Universidade Federal de Santa Catarina.

Além desses dados, o Corpus integra a Antologia de Poemas em Libras[3]. Segundo Machado (2017), uma antologia de línguas de sinais constitui uma forma de registro da cultura surda por meio de produções poéticas. Essa antologia compreende poemas produzidos por surdos com diferentes estilos, objetivando representar a produção poética em Libras, e em breve estará disponível no Corpus de Libras.

O Corpus de Libras compreende também os glossários em Libras de áreas especializadas, disponibilizados por meio de um programa

[3] A Antologia de Poemas compreende dados do Libras Acadêmico e da Tese de Doutorado de Fernanda de Araújo Machado (2017).

desenvolvido pela Universidade Federal de Santa Catarina (ver STUMPF, OLIVEIRA e MIRANDA, 2014).

Novos projetos podem passar a compor o Corpus de Libras indefinidamente, tornando-o mais amplo e variado e compreendendo uma documentação da Libras para ser acessada para diferentes fins (dentre eles, fins históricos), para o ensino e para a pesquisa.

Os dados compreendidos no Corpus de Libras podem incluir também arquivos de transcrição e anotação dos dados. O objetivo destas é permitir a análise linguística sistematizada e, para isso, foram estabelecidas normas para sua realização, objetivando atingir consistência que permita aos pesquisadores compreender os registros realizados, bem como comparar os dados em análise. A seção seguinte apresenta os estudos já realizados sobre transcrições e anotações de línguas de sinais e, posteriormente, discorre sobre o processo de estabelecimento dessas normas para o Corpus de Libras, especialmente para o Inventário Nacional de Libras.

Estudos sobre transcrição de línguas de sinais

A transcrição de dados de corpora de línguas de sinais é necessária por facilitar a análise dos dados. No entanto, o fato de estarmos diante de dados multimodais torna essa tarefa bastante complexa. As línguas de sinais se apresentam na modalidade visual-espacial, com produções corporais envolvendo, normalmente, as mãos, a face e o tronco. Sendo assim, as produções que integram os corpora de línguas de sinais se apresentam em vídeo. Além disso, as línguas de sinais não apresentam um sistema de escrita amplamente disseminado entre os seus usuários. Essas características têm sido alvo de debate entre os pesquisadores de línguas de sinais, uma vez que a transcrição desta depende de uma escrita estabelecida.

Pizzuto e Pientrandrea (2001) discutem sobre esses sistemas de transcrição. O fato de não haver uma escrita consolidada nas línguas de sinais faz com que tenhamos que recorrer ao uso de glosas da língua falada e escrita no respectivo país, pois as línguas de sinais são "orais" e consideradas ágrafas.

Johnston (1991) foi um dos primeiros autores a escrever sobre transcrições de línguas de sinais. Ele trata a transcrição como "um ato de transcrever uma língua que é em si um ato de análise" (JOHNSTON, 1991, p. 4). Pizzuto e Pientrandrea (2001) também mencionam o fato de que o uso de transcrições e anotações são condicionados pelas descrições e análises. Considerando isso, a transcrição exige muita responsabilidade por parte do linguista, pois poderá influenciar as conclusões sobre um fenômeno linguístico. Ainda segundo Johnston, a transcrição codifica a língua com a intenção de torná-la uma unidade analítica. Dessa forma, a transcrição precisa de convenção e padronização. Nessa linha, o autor propõe o uso de glosas da língua escrita de forma padronizada e sistemática, e com consistência nas transcrições de línguas de sinais. As vantagens que o autor apresenta sobre o uso de glosas incluem a simplicidade, a economia de símbolos e a facilidade de acessar a transcrição (pois é mais fácil ler as glosas a ler os sinais). As glosas permitem o acesso aos sinais de forma mais eficiente e rápida. No entanto, as desvantagens em usá-las devem sempre ocupar os linguistas nas análises dos fenômenos linguísticos em línguas de sinais. Johnston menciona como 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. Essas desvantagens precisam, então, ser levadas em conta quando usamos glosas para transcrever sinais, e implicam também em problemas de ordem metodológica.

Para minimizar esses problemas metodológicos, Johnston propõe sistematização das transcrições por meio de convenções. O primeiro passo é definir quais serão as unidades das glosas: o sinal, a sentença

e/ou o enunciado. As glosas sinal por sinal exigem a definição do que compreenderia o início e o final de um sinal. De modo geral, o sinal em fase de transição até a sua forma estável, ou seja, quando está se formando com a sua configuração de mão, locação e movimento, até chegar a sua produção com todos esses fonemas, é considerado o início e o final do tempo para a inserção da glosa.

Apesar de exigir essas decisões de ordem metodológica, a transcrição de cada sinal separadamente é considerada a menos complexa, mas, mesmo assim, há algumas dificuldades impostas pelo uso. Esse tipo de transcrição pode, por exemplo, envolver processos fonológicos e interferir na forma dos sinais. Já a transcrição por sentença é baseada em uma análise sintática. Parte-se do pressuposto da existência de um verbo e de seus argumentos, interno e externo, quando é o caso. Então, estabelece-se a unidade sintática. Esse nível de transcrição exige um conhecimento sintático da língua, que muitas vezes é baseado no conhecimento da escrita da língua falada no país. Isso pode implicar em problemas na análise linguística da língua de sinais em questão. No nível do enunciado, a base é mais semântica. Seria uma opção que também exigiria uma pré-análise envolvendo esse nível de análise.

Tanto Johnston (1991) quanto Pizzuto e Pientrandrea (2001) apontam a segmentação, a representação dos sinais e a dificuldade em reconstruir sinais transcritos por meio de glosas como problemas de ordem teórica e metodológica. Após transcrever os sinais utilizando glosas é muito difícil conseguir reproduzi-los, porque as glosas realmente são limitadas pelas fronteiras estabelecidas pela escrita associada a uma outra língua, que não é visual-espacial. Como Johnston observou, essa é uma das desvantagens do uso de glosas. Assim, o linguista de língua de sinais deve sempre ter em mente que o acesso aos sinais será necessário. A glosa, portanto, tem uma função puramente instrumental.

Um desafio proposto por Nonhebel, Crasborn e van der Kooij (2004) é discutir convenções para transcrição, para criar um banco de dados

comparável para os estudos das línguas de sinais. Assim, as transcrições deveriam ser úteis a quaisquer pesquisadores de línguas de sinais. A proposta dos autores, portanto, é apresentar trilhas de transcrição mais restritas, mas com dados mais abrangentes, que possam ser expandidos pelos pesquisadores de acordo com seus objetivos específicos. Os autores propuseram trilhas de transcrição independentes para a mão direita e para a mão esquerda com os mesmos valores. Vejam que esses autores já estão na era do ELAN, usando um sistema de anotação que permite a inclusão de trilhas para cada aspecto transcrito associado diretamente ao vídeo. A proposta dos autores inclui as seguintes convenções:

a. Glosa utilizando letras maiúsculas para os sinais, utilizando hífen quando for usada mais de uma palavra para identificar um único sinal, por exemplo, DAR-SINAL;
b. (fs-) introduzindo sinais que usaram a soletração: (fs-) NUNCA;
c. IND para sinais de apontação;
d. (2h) para sinais que normalmente são feitos com uma mão e foram feitos com duas mãos: (2h)SONHAR;
e. (1h) para sinais que normalmente são feitos com duas mãos e foram feitos com uma mão: (1h)TRABALHAR;
f. (p-) para sinais polimorfêmicos, com vários componentes. Por exemplo, sinais que envolvem um classificador: (p-)veículo-estacionado-lado-a-lado;
g. (-h) identifica um sinal com final estabilizado/alongado, por exemplo: CORRER(-h);
h. (g-) o "significado" de um gesto é colocado com letras minúsculas e colocado entre aspas: (g-)"bom" e para mãos para cima é usado (g-)pu.

Os autores sugerem que a repetição dos sinais seja anotada por meio de uma trilha independente, indicando a quantidade de vezes que o sinal é repetido. Isso também é sugerido para a marcação espacial. Os autores indicam o uso de trilhas para "cabeça", "posição das sobrancelhas", "olhos", "abertura dos olhos", "direção dos olhos", "forma da boca", "bochechas" e "jogo-papeis". Há também uma trilha para comentários do transcritor, e outra para tradução.

Crasborn (2015) retoma essas questões de ordem metodológica e enfatiza a importância da padronização das convenções das transcrições. O autor chama a atenção para o fato de convenções de transcrições de línguas de sinais não estarem disponíveis publicamente, com algumas exceções (ver Chen Pichler et al., 2010, e Johnston, 2014, como exemplos). O autor observa que publicar as convenções exige também garantir a compreensão por todos do que elas realmente representam, ou seja, elas precisam ser devidamente interpretadas. Além disso, a validação das transcrições também precisa ser incluída nas pesquisas com línguas de sinais.

Em síntese, os aspectos abordados pelos autores a serem considerados sobre as transcrições nas línguas de sinais são os seguintes:

1) Padronização com o objetivo de viabilizar a comparação de dados da língua de sinais coletados e transcritos em diferentes partes do país;

2) Transcrições em diferentes níveis de análise:

a. Transcrições com o mínimo de análise prévia (simplificação), com objetivo de tornar acessível e útil a qualquer pesquisador. A proposta é apresentar uma quantidade de trilhas mais restrita, mas com dados abrangentes, que podem ser expandidos pelo pesquisador de acordo com objetivos específicos (no mesmo sentido de Nonhebel, Crasborn e van der Kooji, 2004). Uso de trilhas mães para a mão direita e para a mão esquerda;

b. Transcrições com análise de acordo com o objetivo da pesquisa, determinadas pelo pesquisador (detalhamento para fins de pesquisa específicos). Possibilidade de criação de novas trilhas mães e de trilhas filhas com tipos linguísticos e vocabulário controlado pré-definidos, de acordo com o objetivo da pesquisa.

3) Uso de glosas para representar sinais manuais da Libras: limitações e avanços;

4) Definição da função da transcrição: servir de referência para organização e compilação de dados para serem usados por ferramentas de busca (a anotação não serve para tentativas de reconstrução dos sinais reais, mas para facilitar a localização de sinais para o desenvolvimento de pesquisas): o que disponibilizar por meio de transcrições?

Considerando tais discussões, apresentamos a seguir as decisões tomadas para a composição das transcrições dos sinais no Corpus de Libras.

Decisões tomadas para a transcrição básica dos sinais do Corpus de Libras

O grupo de pesquisa do Corpus de Libras tomou algumas decisões para a realização das transcrições da Libras no escopo do Inventário Nacional de Libras. Basicamente, as decisões compreendem os seguintes aspectos:

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

2) A anotação apenas por meio de glosas de sinais produzidos, exclusão de informações morfológicas com a utilização do ID

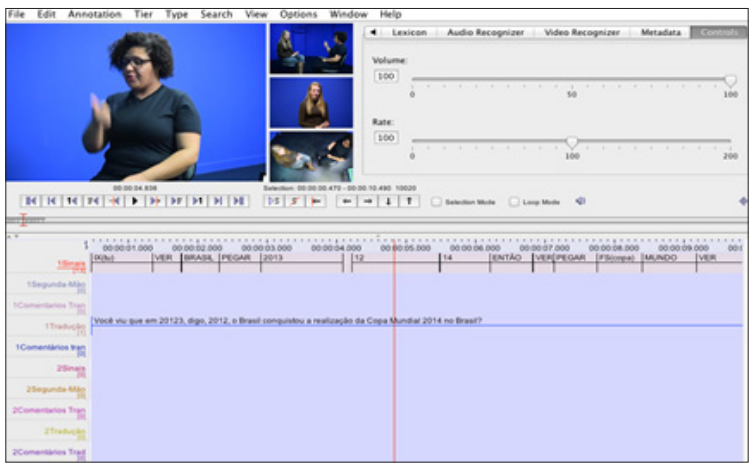
de cada sinal (evita-se o problema em definir o que constituiria a sentença na língua de sinais);

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

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

O Sistema de Anotação *Eudico Annotator* – ELAN (https://tla.mpi.nl/) –, integra as transcrições do Corpus de Libras. O ELAN é um sistema de transcrição que possibilita a visualização de vídeos e a correlação dos dados escritos com os dados nos próprios vídeos. Esse sistema se adequa ao tipo de pesquisa que inclui dados com línguas de sinais. O software permite a criação, edição, visualização e busca de anotações através de dados de vídeo e áudio, e criação de "trilhas" para registro e análises específicas nas duas modalidades de línguas. As convenções para transcrição foram estabelecidas pelo grupo de pesquisa por meio de um manual.

Figura 1: Tela do ELAN com vídeos do Inventário Nacional de Libras



As trilhas de transcrição utilizadas nos arquivos do ELAN são as seguintes:

- 1Sinais D;
- 1Sinais E;
- 1Comentários transcritor;
- 1Tradução PB;
- 1Comentários tradutor;
- 2Sinais D;
- 2Sinais E;
- 2Comentários transcritor;
- 2Tradução PB;
- 2Comentários tradutor.

A numeração inicial indica o sinalizante 1 e o sinalizante 2, pois os dados sempre são coletados em duplas. SinaisD e SinaisE envolvem as trilhas da produção dos sinais, indicando os sinais realizados com a mão direita e a mão esquerda. Todos os sinais produzidos são transcritos um a um, considerando o início do sinal quando o sinalizante inicia a preparação para produzir o sinal, e o final quando inicia a preparação do próximo sinal ou a pausa. Todas as produções dos sinais também são traduzidas para o português. As trilhas de comentários foram incluídas para o registro de observações específicas que os transcritores e os tradutores considerem pertinente anotar.

A transcrição é um processo que demanda um grande investimento de tempo e dedicação, particularmente nas pesquisas com línguas de sinais, que não possuem um sistema de escrita convencional e plenamente adaptado ao computador. Uma estimativa geral relatada em projetos de pesquisa com línguas de sinais é a de uma hora de trabalho de transcrição para cada minuto de gravação[4]. Por esse motivo, e

[4] http://www.sign-lang.uni-hamburg.de/intersign/workshop4/baker/baker.html. Acesso em: 30/06/2016.

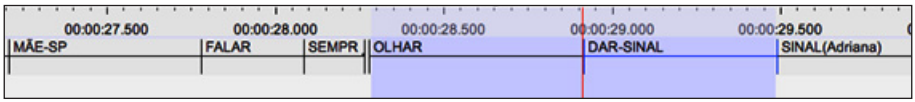
considerando as restrições temporais do Inventário de Libras, foi iniciada a primeira etapa de transcrição do trabalho, envolvendo parte dos dados coletados (em torno de 20 horas)[5]. Nessa primeira etapa, o foco está no desenvolvimento de convenções e critérios para essa transcrição, a partir de amostras dos dados que possam caracterizar elementos do inventário de língua de sinais.

Todas as transcrições precisam passar por um processo de validação. Para isso, membros do projeto com experiência em transcrição realizam uma segunda transcrição em amostras estatisticamente significativas dos dados coletados, com fins de comparação com as transcrições originais. Esse processo deve ser realizado periodicamente, a fim de avaliar o processo de transcrição e introduzir ajustes quando necessário. Além disso, contamos com um pesquisador encarregado de revisar a transcrição original em busca de inconsistências com relação às convenções de anotação desenvolvidas no projeto. Todas as convenções de transcrição do projeto ficam disponibilizadas permanentemente a todos os pesquisadores no Portal de Libras (www.libras.ufsc.br).

A anotação dos sinais por meio de glosas com palavras da Língua Portuguesa foi definida como forma das transcrições dos textos em Libras com a utilização do ELAN. As glosas ainda são consideradas mais simples para transcrever sinais, assim como indicado por outros pesquisadores já mencionados anteriormente (JOHNSTON, 1991; PIZZUTO e PIENTRANDREA, 2001; NONHEBEL, CRASBORN e VAN DER KOOJI, 2004, por exemplo). As glosas usadas compreendem o sinal em si sem as informações quanto a mudanças morfológicas flexionais. Essa decisão foi determinada para tornar o sistema de transcrição mais simples, exigir menos análise prévia de cada sinal por parte do transcritor, ou seja, cada sinal corresponde a uma glosa pré-definida sem marcações flexionais.

[5] Contamos com uma média de 5 bolsistas para realizar transcrição de dados. Em 2014-2016 contamos com cinco bolsistas transcritores de iniciação científica: Marcos Marquioto (CNPQ), Bianca Gomes (CNPQ), Miriam Royer (CNPQ), Edinata Camargo (voluntário) e Harrison Adams (voluntário).

Figura 2: Exemplo de sinais com IDs sem marcação flexional



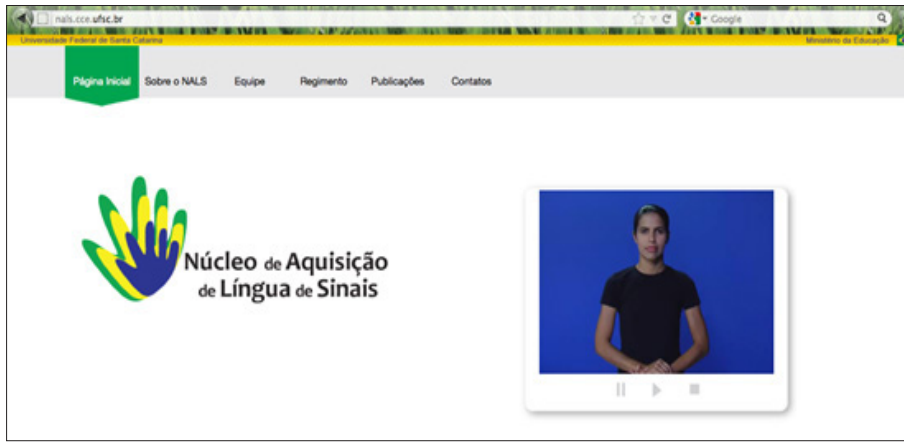
Os verbos OLHAR e DAR-SINAL são flexionados em Libras, mas são apresentados por meio de IDs na forma infinitiva, independentemente da forma que o sinal apresenta com ou sem flexão. Se o pesquisador quiser analisar o sistema flexional dos verbos na Libras, terá que criar trilhas específicas para criar anotações de análise de cada ocorrência verbal e, então, identificar o padrão. Posteriormente, voltaremos às anotações decorrentes de análises específicas.

Para usar glosas, criamos o Identificador de Sinais, que é um banco de dados de sinais que serve como um instrumento metodológico para nomear sinais por meio de glosas. Ao usarmos os identificadores de sinais em nossas transcrições, passamos a ter condições de torná-las mais eficientes, devido aos sistemas de buscas existentes no sistema de transcrição que usamos (*Eudico Annotator* – ELAN). Isso tem facilitado imensamente as pesquisas em andamento, bem como projetos futuros que envolvam análises de produção em sinais. O Identificador de Sinais está disponível de forma aberta e gratuita para todos os interessados em utilizá-lo e alimentá-lo como fonte de pesquisa, na página: http://www.idsinais.libras.ufsc.br. A proposição de desenvolver um Identificador de Sinais foi motivada pelo uso que temos feito das transcrições. Sempre nos deparávamos com as inconsistências das transcrições realizadas por nós mesmos ou por bolsistas que participaram de nossas pesquisas. O contexto de realização dos sinais, bem como as interpretações feitas por cada transcritor tornavam as transcrições inconsistentes. Isso também foi observado por outros pesquisadores de línguas de sinais, no mundo inteiro. Em congressos e seminários, essa questão era recorrente quando se discutiam questões metodológicas que norteavam as pesquisas

com línguas de sinais. Além disso, como temos bolsistas que vêm de diferentes partes do Brasil, variantes das línguas faladas acabavam sendo usadas para nomear sinais comuns e recorrentes em nossos documentos. Assim, chegamos em um estágio que sentimos a necessidade de contar com algum registro para as glosas atribuídas aos sinais. Começamos registrando em um documento comum os sinais e suas respectivas glosas. Essas glosas eram discutidas pelo grupo de pesquisa, com base nos dados das próprias crianças. Ao definir a glosa para um determinado sinal, tirávamos uma foto do vídeo da própria criança para identificar o sinal e registrávamos o nome eleito para ser utilizado pelos transcritores. Aos poucos, a lista de sinais com glosas começou a ficar imensa. Os transcritores tinham dificuldade em localizar um determinado sinal para passarem a utilizar em suas transcrições. A quantidade de dados exigia algum tipo de sistematização desse processo para facilitar a organização e sistematicidade dos termos usados para identificar sinais. Aí surgiu, então, a necessidade do Identificador de Sinais.

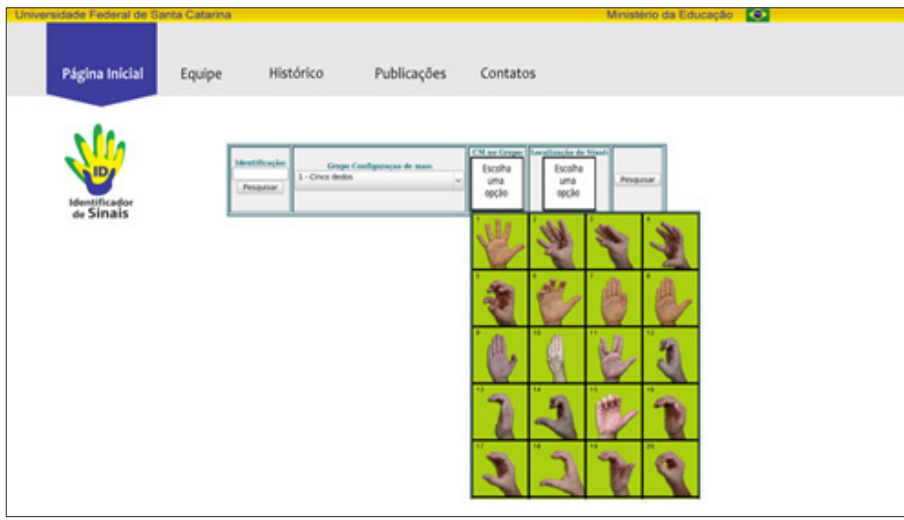
O Identificador de Sinais – ID – é uma ferramenta que disponibiliza os nomes dados aos sinais para as glosas utilizadas nos sistemas de transcrição, bem como a respectiva escrita desse sinal, utilizando a escrita de sinais. Atualmente, o ID compreende 4.000 sinais, que foram levantados por meio de reuniões periódicas realizadas com a equipe de pesquisa do Corpus de Libras. O grupo se reúne e debate sobre os sinais que surgem nos vídeos que estão sendo descritos e "batiza" os sinais que ainda não foram batizados. Os sinais batizados com um ID são imediatamente incorporados no sistema de identificadores de sinais. Atualmente, o ID também pode ser indicado por meio do grupo do Inventário de Libras no WhatsApp. Os transcritores postam trechos dos vídeos originais e debatemos, por meio do WhatsApp, o melhor ID a identificá-lo. Essa metodologia de definição dos IDs tem sido bastante eficiente. Na sequência, o ID é imediatamente incorporado ao Identificador de Sinais. Com os IDs dos sinais, os transcritores realizam a transcrição com mais segurança.

Figura 3: Página inicial do Identificador de Sinais – www.idsinais.libras.ufsc.br



O Identificador de Sinais apresenta um sistema de busca por sinais ou por nomes de sinais. O sistema de busca por sinais tem dois filtros de pesquisa: a configuração inicial do sinal e a localização do sinal (figura 4).

Figura 4: Busca pela configuração de mão



O resultado da busca apresenta as opções dos sinais, conforme ilustrado na figura 5. Cada nome identificado para um sinal conta com uma tradução para o português e para o inglês (figura 6).

Figura 5: Resultados da busca

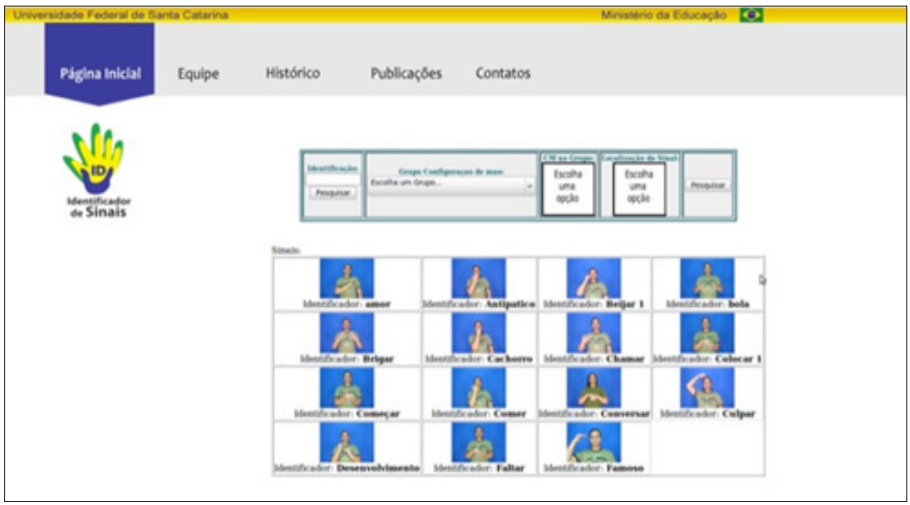


Figura 6: O sinal identificado e suas traduções

A tradução para o português é muito importante, pois nem sempre o nome dado ao sinal corresponde a palavra em português usada em um determinado contexto. Por exemplo, o sinal COMER pode ser traduzido para o português como "comer" ou "comida", mas o nome mantém-se COMER. Na transcrição, tanto no contexto de "comer" quanto no contexto em que é usado o termo "comida", a glosa usada será COMER. Na tradução da transcrição para o português é que você vai utilizar o termo adequado ao contexto. Veja o exemplo a seguir:

Transcrição usando IDs:

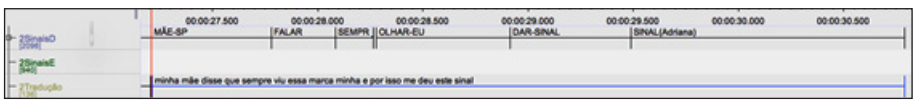
FRUTAS, VERDURAS, PÃO TERMINAR. PRECISAR IR MERCADO COMPRAR COMER. IX(si) SABER IX(João) PRECISAR MAIS COMER, PORQUE COMER COMPLETO.

Tradução para o português:

As frutas, as verduras e o pão terminaram. É preciso ir ao supermercado comprar comida. Eu sei que o João vai precisar de mais comida, pois ele comeu tudo que tinha.

Os tradutores acompanham os vídeos e realizam a tradução considerando o texto em Libras. Eles traduzem com base na semelhança interpretativa, no sentido de Rodrigues (2014). Eles assistem ao vídeo em Libras e realizam a tradução com base nos sentidos em forma de enunciados. Veja o exemplo a seguir:

Figura 7: Exemplo de trecho de tradução de um enunciado

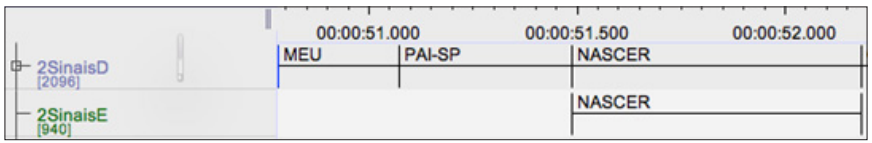


O transcritor recorre ao Identificador de Sinais para ver como nomear o sinal para "comer" e "comida", pois, na Libras, os dois itens são homônimos. Assim, ele vai se certificar da glosa eleita para aquele sinal e das suas possíveis traduções para o português. A tradução para o inglês também fica disponível para outros pesquisadores de outras línguas

de sinais, bem como para tradução de glosas em artigos internacionais publicados em inglês.

Cada sinal é transcrito individualmente; na trilha da mão direita, se for produzido com a mão direita, e na trilha da mão esquerda, se for produzido com a mão esquerda. Se for produzido com ambas as mãos, o ID é repetido em ambas as trilhas. Veja o exemplo a seguir:

Figura 8: Exemplo de transcrição com sinal produzido com as duas mãos – SinalD (mão direita) e SinalE (mão esquerda).



Os dados são coletados por meio de câmeras, podendo compreender diferentes perspectivas da interação em sinais. No caso do Inventário Nacional de Libras, estamos usando 4 câmeras para coletar as produções de surdos agrupados de dois em dois. A primeira câmera foca no sinalizante à direita, a segunda no sinalizante à esquerda, a terceira nos dois sinalizantes de frente e a quarta nos dois sinalizantes de cima para baixo. Veja a figura com as quatro perspectivas.

Figura 9: Perspectivas das filmagens do Inventário Nacional de Libras

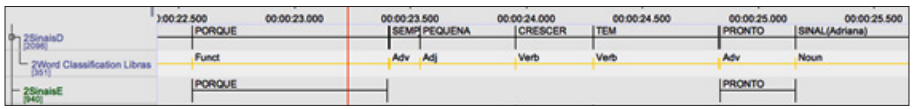
As quatro perspectivas precisam ser sincronizadas no tempo para serem transcritas simultaneamente, considerando cada participante da conversa. Os transcritores podem escolher as perspectivas que irão usar de acordo com o participante que estiverem transcrevendo. Alguns transcritores preferem ter todas as perspectivas disponíveis, pois consideram que isso torna os sinais mais claros e facilita a transcrição. Outros preferem manter apenas o vídeo que tem a perspectiva do sinalizante que estejam transcrevendo, junto com o terceiro vídeo, que inclui o seu interlocutor. O vídeo de cima para baixo foi considerado importante por vários transcritores, por facilitar a visualização da dimensão do sinal, que não fica completamente visível nos vídeos de frente. A qualidade dos vídeos também é importante para a transcrição, pois facilita a visualização dos sinais, especialmente quando a marcação de tempo é mais lenta. Os transcritores recorrem ao controle de tempo para visualizar os sinais que são executados muito rapidamente e com a transição afetada por processos fonológicos. O vídeo apresentado mais lentamente possibilita identificar o sinal nesses contextos, permitindo a transcrição.

Com base nas decisões tomadas quanto às transcrições, constituiu-se o Manual de Transcrição da Libras. Este é atualizado sistematicamente, apresentando versões de acordo com suas mudanças e indicadas pela data dos ajustes realizados. Os ajustes são feitos na medida em que encontramos novos aspectos que requerem ajustes. A versão atual é de 2013, e ainda não sofreu nenhum ajuste desde então. Provavelmente, chegamos em uma versão mais consistente, depois de vários ajustes feitos desde a sua primeira versão estabelecida para os dados de aquisição da linguagem, em 2002. O Manual de Transcrição do Corpus de Libras foi ajustado para compreender as transcrições básicas dos sinais. Isso também justifica a consolidação desse manual que está disponível na página do Corpus de Libras para consulta.

Anotação de dados para fins de pesquisa dos dados do Corpus de Libras

No caso de anotações para análises que compreendem aspectos linguísticos que requerem trilhas específicas de análise, o manual não apresenta decisões, pois tais especificidades dependem de cada projeto de pesquisa e de seus objetivos. A orientação que temos dado aos pesquisadores é de partirem das transcrições básicas que compreendem os sinais produzidos pelas mãos direita e esquerda, e estabelecerem os procedimentos para a anotação de aspectos específicos de acordo com os seus objetivos de pesquisa. A exemplo, nossas pesquisas já estabeleceram objetivos de análise das classes de palavras compreendidas nos textos em Libras. Para isso, foi criada uma trilha específica para Classes de Palavras com vocabulário controlado, incluindo cada classe (Nome, Verbo, Adjetivo, Advérbio, Pronomes etc.).

Figura 10: Exemplo de anotação das classes de palavras



Essas anotações foram marcadas a partir dos sinais já transcritos previamente na transcrição básica. Então, o pesquisador anotou cada classe para cada sinal. Após esse nível de anotação, é possível estabelecer novos níveis de análise, por meio de novas anotações. Por exemplo, se o objetivo for analisar os elementos recorrentes que compreendem a forma dos sinais de cada classe de palavras na Libras, podem ser criadas novas trilhas específicas para cada classe de palavra com vocabulário controlado específico, tais como a especificação do tipo de movimento, das marcações não-manuais, o uso de uma ou duas mãos, e assim por diante. A transcrição, como disse Johnston (1991), apresenta uma análise em si mesma. No momento em que o pesquisador define o detalhamento

das anotações considerando seus objetivos de pesquisa, as informações anotadas já estarão organizadas para análise pretendida. Utilizando o ELAN, o pesquisador vai, então, estabelecer as trilhas que serão inseridas a partir das trilhas básicas do Corpus de Libras.

Considerações finais

Na linha dos demais pesquisadores, nós concordamos que as transcrições de línguas de sinais utilizando IDs (glosas com palavras da Língua Portuguesa) já incluem análises prévias e podem influenciar a análise dos fenômenos linguísticos analisados compreendidos em um texto em língua de sinais. Assim, ao longo dos anos, fomos aperfeiçoando a metodologia para realizar as transcrições da Libras. Com o início do Corpus de Libras, concebemos um Manual de Transcrição da Libras que chegou à sua forma atual já em 2013. Esse manual apresenta as convenções que estamos usando para nortear as transcrições da Libras no escopo do Corpus de Libras. Preocupados com a sistematização e consistência das transcrições, nos ocupamos de criar o Identificador de Sinais, um software que disponibiliza um banco de sinais com suas identidades nomeadas por meio de uma única glosa, representada por uma palavra da Língua Portuguesa. Cada sinal passa a ter uma ID (uma identidade, uma única glosa) para ser acessada pelos transcritores no momento da transcrição. Esse banco de sinais tem sido alimentado sistematicamente a partir das próprias produções em Libras que compõem o Corpus de Libras, por meio de diferentes projetos. A transcrição é uma etapa complexa das pesquisas com línguas de sinais e, por isso, exige o desenvolvimento de estratégias metodológicas consistentes, padronizadas e sistemáticas para garantir a possibilidade de acesso aos dados por diferentes pesquisadores. Esses avanços têm sido possíveis também por meio da tecnologia, que permite a organização de softwares que favoreceram a compilação e a

sistematização dos dados. Além disso, o fato de contarmos com o ELAN para realizar as transcrições tem sido significativo para as pesquisas com a Libras. O Corpus de Libras está, portanto, acompanhando essas evoluções tecnológicas e metodológicas, no sentido de garantir a qualidade dos dados disponibilizados aos pesquisadores da Libras.

Agradecimentos

Agradeço aos pesquisadores assistentes Deonísio Schmitt, Juliana Tasca Lohn, Aline Lemos Pizzio e Bruna Crescêncio Neves. Agradeço aos pesquisadores colaboradores Diane Lillo-Martin, Deborah Chen Pichler, Christian Rathmann, Jair Silva e Rachel Sutton-Spence. Também agradeço a todos os bolsistas de iniciação científica que contribuíram para a consolidação das convenções do Manual de Transcrição da Libras, em especial, Miriam Royer, Bianca Gomes Sena, Harrison Adams, Edinata Camargo, Luana Marquezi e Karina Christmann. Agradeço ao técnico de informática Ramon Dutra Miranda e ao apoio técnico Roberto Dutra Vargas. Enfim, agradeço ao CNPQ pelo financiamento das pesquisas.

Referências

CHEN PICHLER, D.; HOCHGESANG, J.; LILLO-MARTIN, D.; QUADROS, R. M. de. **Conventions for Sign and Speech Transcription in Child Bimodal Bilingual Corpora**. Language, Interaction and Acquisition/Language, Interaction et Acquisition 1(1), 11-40. 2010.

CRASBORN, O. A. Transcription and Notation Methods. In **Research Methods in Sign Language Studies***:* A Practical Guide, First Edition. Edited by Eleni Orfanidou, Bencie Woll, and Gary Morgan. John Wiley & Sons, Inc. 2015.

JOHNSTON, T. Transcription and glossing of sign language texts: examples from AUSLAN (Australian Sign Language). In **International Journal of Sign Linguistics**. Multilingual Matters. Vol.2:1. 1991.

_____. **Auslan corpus annotation guidelines**. Manuscript, Macquarie University, Sydney. Accessed October 29, 2014. Disponível em: <http://new.auslan.org.au/about/annotations/>

MACHADO, F. de A. **Antologia de Poemas em Libras**. Tese (Doutorado). Programa de Pós-Graduação em Estudos da Tradução. Universidade Federal de Santa Catarina, 2017.

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.

NONHEBEL, A.; CRASBORN, O.; van der KOOIJ, E. **Sign language transcription conventions for the ECHO Project**: BSL and NGT mouth annotations. Disponível em: https://www.researchgate.net/publication/237538700_Sign_language_transcription_conventions_for_the_ECHO_Project. 2004.

PIZZUTO, E.; PIENTRANDREA, P. The notation of signed texts: open questions and indications for further research. In **Sign Language & Linguisitcs**. John Benhamins Publishing Company. 43/2. 29-45. 2001.

QUADROS, R. M. de. Documentação da Libras. (2016) In **Seminário Ibero-Americano de Diversidade Linguística**, 2014, Foz do Iguaçu. Brasília: IPHAN – Ministério da Cultura. v. 1. p. 157-174.

RODRIGUES, C. A busca por semelhança interpretativa no processo de interpretação simultânea para a língua de sinais. In Quadros, R. M. de;

Weininger, M. (Org.). **Série de Estudos de Línguas de Sinais**. Volume 3. 35-69. (2014)

STUMPF, M.; OLIVEIRA, J.; MIRANDA, R. D. Glossário Letras Libras, trajetória dos sinalários no curso - como os sinais passam a existir. In Quadros, R. M. De (Org.) **Letras Libras: Ontem, Hoje e Amanhã**. Florianópolis: Editora da UFSC. (2014).