DOI: 10.18468/rbli.2018v1n1.p23-45

Documentação linguística na Amazônia: abordagens e métodos[1]

Language Documentation in the Amazon: approachs and methods

Glauber Romling da Silva
Licenciatura Intercultural Indígena
Universidade Federal do Amapá – Campus Binacional de Oiapoque

RESUMO: Este artigo tem por objetivo apresentar um panorama da área da Documentação Linguística a partir de experiências recentes na Amazônia brasileira. Primeiramente, traçamos um breve histórico da área. A partir disso, refletimos sobre o seu atual estado da arte e extraímos alguns conceitos básicos que devem nortear um projeto de documentação linguística de êxito (a dupla necessidade, o triplo impacto e os quatro princípios de um corpus). Adiante, apresentamos os resultados de um projeto de documentação desenvolvido com a língua Paresi-Haliti, que exemplificam esses conceitos, e descrevemos os pontos importantes de sua execução. Nesse breve estudo de caso, destacamos a abordagem participativa e autônoma, que busca envolver o máximo de agentes das comunidades afetadas como protagonistas no processo de documentação. Essa abordagem é ajustada a um fluxo de trabalho eficiente, que envolve a utilização de ferramentas digitais de documentação linguística. Para isso, projetos dessa natureza devem inserir em suas agendas o treinamento e a formação de pesquisadores indígenas em métodos e técnicas de documentação linguística, para que, assim, possam prosseguir com ações de documentação após a finalização dos projetos. Concluímos com uma reflexão sobre os principais desafios de base que envolvem a Documentação Linguística, em três eixos: (i) diretrizes éticas sobre materiais culturalmente relevantes, (ii) como torná-la, de fato, emancipadora e (iii) como tornar o processo de documentação tecnicamente eficiente.
Palavras-chaves: Documentação linguística. Paresi-Haliti. Ferramentas digitais.

ABSTRACT: This article's goal is to present an overview of the Language Documentation field from recent experiences in the brazilian Amazon. First of all, we draw a brief history of the field. From that, we reflect upon its current state-of-the-art and extract some basic concepts which should guide a successful language documentation project (the double need, the triple impact and the four principles of a corpus). Forward, we present the results of a documentation project developed with the Paresi-Haliti language, exemplifying these concepts, and describe the important points of its execution. In this brief case study, we highlight the participatory and autonomous approach, which seeks to involve as many agents from the affected communities as protagonists in the documentation process. This approach is adjusted to an efficient workflow, involving the utilization of language documentation digital tools. To that, projects of this nature must insert in their agendas the training and education of indigenous research-

[1] Este trabalho foi financiado por um Field Trip Grant do Hans Rausing ELDP/SOAS de 2008 a 2013, pelo Projeto de Documentação de Línguas Indígenas do Museu do Índio/FUNAI/UNESCO de 2009 a 2013 e por uma bolsa de doutorado concedida pelo CNPq de 2010 a 2013.

ers in language documentation methods and techniques, so that they can proceed with documentation actions after the completion of the projects. We conclude, with an observation over the main base challenges involving the Language Documentation, in three axis: (i) ethical guidelines on culturally relevant materials, (ii) how to make it, in fact, emancipatory and (iii) how to make the documentation process technically efficient.
Keywords: Language documentation. Paresi-Haliti. Digitals tools.

1. Introdução

Este trabalho tem por objetivo discutir o que é necessário para um projeto de documentação cumprir eficientemente suas demandas científicas e humanas. No campo científico, a área da documentação linguística tem um impacto em termos empíricos, analíticos e metodológicos; no campo humano, colabora para a dinamização, consolidação e efetivação de políticas linguísticas. Para a efetivação desse poder transformador da documentação linguística, a evolução no acesso a programas computacionais de documentação linguística foi essencial para dar um salto na capacidade de trabalho de pequenos times, mas ainda torna o processo muito consumidor de tempo. Apresentamos os conceitos de dupla-necessidade, que guia as finalidades de um projeto de documentação; triplo impacto, que norteia seus objetivos específicos, e dos quatro princípios, que funda as bases metodológicas na construção de *corpora* eficientes. Para isso, tomamos como exemplos projetos de documentação empreendidos na Amazônia brasileira. Buscamos apresentar uma contribuição para o campo da Documentação Linguística, tendo como pano de fundo o estabelecimento de limites que nos levam a crer que estamos na transição para uma nova era da documentação definida pela radicalização do conceito de autonomia.

1.1. Breve histórico da Documentação Linguística

A **documentação linguística**, como área específica, **define-se como** a compilação e preservação de dados linguísticos primários e secundários e as interfaces entre esses dados e os vários tipos de análises linguísticas possíveis baseadas nesses dados (Gippert *et alii*, 2006). Desde o início dos anos 90, assume-se que, se nada fosse feito, nos próximos decênios, a diversidade linguística cairia vertiginosamente. Essa projeção, caso concretizada, teria fortes consequências para a ciência linguística e para os falantes dessas línguas (Hale *et alii*, 1992; Krauss, 1992).

A história da Documentação Linguística divide-se em três grandes fases: a era pré-90, a era pós-90 e a era atual. **A era pré-90** é composta pelos herdeiros da tradição do tripé de Franz Boas, em que uma boa documentação deveria, essencialmente, gerar três produtos: um bom léxico, um esquete gramatical e uma compilação de textos anotada (Jakobson, 1944). A **era pós-90** define-se pela inserção de alguns fatores que dinamizaram a tarefa de documentar línguas em perigo: (i) a abordagem participativa com as comunidades envolvidas e (ii) o acesso a novas tecnologias, como os programas de documentação

linguística. Essa abordagem para a **autonomia** prima pela **participação efetiva** das comunidades no processo de documentação. As comunidades envolvidas têm a chance de participar na modelagem dos projetos, estabelecendo prioridades e demandando contrapartidas claras dos pesquisadores, como a produção de materiais didáticos ou a criação de centros de documentação nas aldeias. Através do treinamento em métodos e técnicas de documentação linguística, membros das comunidades envolvidas participam do processo de documentação de maneira ativa. A abordagem participativa é uma consequência lógica e imediata da grande tomada de consciência de nações minoritárias quanto ao seu papel e posição na relação com pressões culturais majoritárias. O acesso a novas tecnologias, como programas de documentação para anotação linguística (ELAN[2]), visualização de dados linguísticos (SHOEBOX, TOOLBOX, FLEX[3]) e confecção de *metadata* (Arbil) deram poder às equipes de projetos de documentação de trabalharem com uma grande quantidade de dados de maneira mais rápida, econômica e eficiente.

A fronteira cronológica que delimita a era pós-90 e a atual ainda está sendo construída. Podemos definir como peça-chave para postularmos essa transição a inclusão de um desafio preponderante, que lhes confira mais autonomia: **como a comunidade científica e as comunidades linguísticas envolvidas podem potencializar a criação de novos corpora anotados?**

Para definirmos isso, situamos nossa abordagem para a Documentação Linguística em alguns conceitos. A dupla necessidade (o científico e o humano) subjacente às motivações em documentar uma língua, o triplo impacto (a empiria, análise e o método) referente aos resultados esperados de um trabalho dessa natureza e os quatro princípios (diversidade, acessibilidade, expansibilidade e flexibilidade) que devem servir de guia para a construção de um *corpus* elevante. Neste trabalho apresentaremos uma experiência concreta em projeto de documentação empreendido em comunidades indígenas na Amazônia e refletiremos sobre seus potenciais científicos e humanos, e sobre seus desafios no atual estado da arte da área.

1.2. A dupla necessidade: o científico e o humano

A Documentação Linguística cresceu exponencialmente nas últimas décadas por conta de alguns fatores catalisadores. O acesso a novas tecnologias na área e a emergência de programas computacionais de documentação com interfaces cada vez mais amigáveis fundaram os pilares de uma nova era na Linguística. Predecessora ao ápice desse dois motores, estava a preocupação com as línguas em perigo ou em iminência de extinção. Em paralelo, residia o impulso científico de catalogar, registrar e preservar dados primários de maneira sistemática e científica. Documentar línguas em iminência de extinção significa

[2] Lausberg, H., & Sloetjes, H. (2009). Disponível para download no site **Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands** http://tla.mpi.nl/tools/tla-tools/elan/download/.

[3] Disponível em sil.org

salvar memórias, culturas e símbolos emblemáticos que dão identidade. Preocupar-se com a qualidade e a precisão dos registros e de suas anotações, através de metodologia própria e com objetivos bem definidos, contribui para o acesso da Linguística como ciência a novas questões, antes veladas pela incapacidade de construir *corpora* diversos e massivos.

A emergência do acesso a novas tecnologias de documentação linguística mais necessidades de caráter **científico** e **humano** fundaram um novo campo na Linguística. Necessidades **científicas** são as que versam sobre os meios e métodos de construir *corpora* organizados e úteis para diversos campos científicos que necessitam de dados linguísticos para seus processos de argumentação; já necessidades **humanas** são aquelas que partem das comunidades que recebem equipes de documentação linguística, que são, em geral, minoritárias e com acesso frágil às manufaturas do conhecimento, a saber, ortografias consolidadas, livros didáticos, livros de narrativas tradicionais, etc. Defendemos neste trabalho que, para suprir essas duas necessidades principais, um *corpus* deve seguir quatro princípios: ser diverso, acessível, expansível e flexível.

1.3. O triplo impacto: a empiria, a análise, o método

Não basta ter um repositório tipologicamente bem construído, ele deve poder ser usado de maneira eficiente para fins científicos e humanos. Quanto à sua dimensão **empírica**, a Documentação Linguística pode revelar maior diversidade de línguas. Na dimensão **analítica**, a partir do acesso a uma maior diversidade tipológica de línguas, ganha-se mais poder de falsificar hipóteses antes apenas abstratas. E, por fim, **metodologicamente**, buscam-se metodologias mais apuradas que possam ser replicáveis e controláveis (Gomes, 2015; Lima, 2015; Davis *et alii*, 2014).

Apresentamos neste trabalho uma reflexão sobre essa mudança e os rumos que a área da documentação vem tomando nos últimos anos. Para a área avançar no eixo humano, plataformas computacionais de documentação com interface multiusuário que possibilitem a participação maior do consultor ao processo de construção do *corpus* podem ser a resposta para a superação de diversas barreiras que ainda tornam a documentação uma tarefa árdua e muito consumidora de tempo.

1.4. Os quatro princípios de um corpus: diversidade, acessibilidade, expansibilidade e flexibilidade

Um *corpus* linguístico deve ter um formato que atenda às suas funções. Como, em geral, desejam-se *corpora* mais flexíveis que atendam a públicos diferentes (linguistas, antropólogos, arqueólogos, comunidades indígenas, leigos, etc.) com propósitos distintos (pesquisa científica, pesquisa escolar, publicações na imprensa, exposições, mera curiosidade, etc.), define-se que *corpora* bem construídos devem ter qualidades que garantam essa flexibilidade. Um *corpus* deve ser diversificado, acessível, expansível e flexível em seu formato e função. Para que isso aconteça, alguns cuidados são necessários.

Tabela 1: Os quatro princípios

CORPORA	Função	Formato
Diversidade	Ter diferentes gêneros	Mídias apropriadas para situações específicas de registro
Acessibilidade	Ser organizado otimamente	Metadata intuitivos e sistemas de rotulagem
Expansibilidade	Permitir expansão infinita	Um sistema coerente de categorização
Flexibilidade	Ter estruturas atualizáveis	Formatos básicos e livres para mídias

Para construir um *corpus* **diverso**, o linguista e sua equipe devem estar prontos para situações distintas. A racionalização dos recursos materiais de registro (gravadores e câmeras) em relação ao seu uso preferencial em situações distintas é o ponto-chave para um projeto de documentação que almeja construir um *corpus* diverso.

Tabela 2: Registros devem ser diversos

SESSÕES	Tipo de planejamento	Local
Áudio	Planejadas	Internas ou externos a habitações
Video	Não planejadas	Externas

Em situação de campo, o uso do recursos materiais deve levar em conta o tipo de planejamento de sua sessão e o local em que ela ocorre. Para sessões planejadas cujo foco não necessite de documentação visual e que ocorram em ambientes internos ou externos, registros apenas em áudio, com microfones XLR, que necessitam ser acoplados à cabeça ou à roupa do consultor, são suficientes. Já a documentação em vídeo não necessita de preparação prévia do equipamento e pode estar sempre a mão para o registro de sessões não planejadas, por isso, é a mais indicada para sessões incidentais que ocorram em ambientes externos.

Para construir um corpus **acessível**, o linguista deve fazer *metadata* amigáveis e relevantes. Para isso ele precisa ser específico e intuitivo. Devem-se escolher estruturas que cubram especificidades de seu acervo em uma rotulagem com informações básicas, condensadas em rótulos enxutos. Códigos alfanuméricos que incluam o nome da língua, o coletor, a mídia de coleta original, a data, a sequência da sessão no dia e uma pequena descrição livre da sessão servem como exemplo:

(1) PAJSWAV2009093001cura ou PA_JS_WAV_2009_09_30_01_cura
Lê-se: sessão da língua Paresi-Haliti (PA), gravada por José da Silva (JS), em formato WAV, em 2009, no mês 09, dia 30, sequência 01 do dia, cujo conteúdo refere-se a uma cura.

Para construir um corpus **expansível**, o linguista e sua equipe deve preocupar-se com a manutenção de seu *corpus* e seu gerenciamento. Nesse sentido, devem-se evitar repetição de rótulos e ser multimídia. A adoção de um esquema de rotulagem como em (1) garante que rótulos não se repitam e abordagens multimídia facilitam a expansão, uma vez que exploram mais possibilidades de registros em situações diferentes.

Construir um *corpus* **flexível** consiste em se preocupar com as imperfeições que seu *corpus* revelará com o tempo. Para isso, formatos básicos facilitam a migração para formatos novos, conforme a tecnologia avança, pois são mais atualizáveis e legíveis. Formatos de *software* livre ou básicos, com extensões de arquivo *txt*, oferecem total poder de migração dos *corpora* para novos formatos conforme as tecnologias de armazenamento digital atualizam-se. A possibilidade de migração de bases de dados para novas tecnologias é fundamental para a durabilidade e continuidade de um *corpus.* Para a ciência isso é bom, pois conceitos básicos tornam mais fácil a adaptação de *corpora* para outros sistemas, como sistemas de anotação e *parsing* automáticos. Para as comunidades indígenas envolvidas, conceitos básicos tornam mais fácil a migração de *corpora* disponíveis para plataformas mais amigáveis para usuários leigos.

2. Resultados do Projeto de Documentação Paresi-Haliti

Através de um estudo de caso, a documentação da língua *Paresi-Haliti*, apresentamos e discutimos abordagens possíveis que podem, ao mesmo tempo, atender aos anseios culturais da documentação, contrapartida por excelência para as comunidades envolvidas, e fomentar novas possibilidades de investigação científica provenientes do acesso a *corpora* multimídia. Desse modo, serão pontuadas questões imanentes dos conceitos da dupla necessidade, que funda a origem dos projetos de documentação, o triplo impacto desencadeado e os quatro princípios que norteiam a construção de *corpora* linguísticos.

Baseamo-nos, em suas esferas de concepção, desenvolvimento e resultados, em dois projetos de documentação: *Documentation of the Paresi-Haliti Language (Arawak)*, que teve suporte do *Endangered Languages Documentation Programee* da *School of Oriental and African Studies* (*University of London*)[4], através de um *Field Trip Grant,* e Documentação da Língua Paresi-Haliti: uma Língua Arawak do Sul[5], projeto institucional que operou no âmbito do Projeto de Documentação de Línguas Indígenas (PRODOCLIN) do Museu do Índio da Fundação Nacional do Índio (FUNAI) e que teve como parceiro a *United Nations Educational, Scientifc and Cultural Organization* (UNESCO) e a Fundação Banco do Brasil.

2.1. Língua, povo e território

O Paresi-Haliti é uma língua Arawak (ou Aruák), ramo Arawak do sul (sub-ramo Paresi-Saraveka) na definição clássica de Aikhenvald (1999) (Moore *et alii* (2009) e Ramirez (2001) também fazem um levantamento de sua situação). De acordo com dados do Instituto Sócio-Ambiental[6], os Paresi-Haliti somavam 2.005 indivíduos em 2008. Dados do CCGE-

[4] O site do projeto pode ser visitado no link do ELDP SOAS em http://elar.soas.ac.uk/deposit/0085

[5] O site do projeto pode ser visitado no link do PRODOCLIN do Museu do Índio http://prodoclin.museudoindio.gov.br/index.php/etnias/haliti-paresi/povo

[6] http://pib.socioambiental.org/pt/povo/paresi/2031

O-FUNAI afirmam que atualmente essa população está distribuída em 7 áreas indígenas concentradas no estado de Mato Grosso, Brasil. Em informações coletadas por Silva (2013a), juntamente a membros da comunidade, estima-se que, atualmente, o contingente populacional se distribui em 50 aldeias, que estão em constante subdivisão.

Os Paresi-Haliti autodenominam-se *haliti*, que significa 'gente, povo'. Na literatura multidisciplinar existente, encontram-se diversas grafias diferentes para a designação do povo e da língua: Pareci, Parecis, Paresí Paressí, Ariti e Aliti. Atualmente, os membros da comunidade preferem ser chamados de Paresi-Haliti, ou simplesmente Paresi, com a grafia 's' e sem acento agudo no 'i'. Por isso, utilizaremos neste trabalho essas formas.

Esta pesquisa foi realizada na região da Terra Indígena do Rio Formoso, município de Tangará da Serra – MT. Atualmente, a TI Rio Formoso têm as seguintes aldeias: JM, Cachoeirinha, Jatobá, Formoso, Formoso II e Queimada. Empreendemos nossa pesquisa, principalmente, nas aldeias Formoso e Cachoeirinha, com passagens pelas aldeias JM e Queimada.

A aldeia Rio Formoso é a aldeia mais populosa da Terra Indígena do Formoso que, originalmente, tem como núcleo a aldeia Queimada. A expansão para o corredor do Formoso, com sua aldeia principal homônima e pequenas aldeias, deu-se no final da década de 70 e início da década de 80, quando o cacique local, que faleceu em outubro de 2011 com cerca de 95-100 anos, abriu os caminhos para esses lados. Gradativamente, famílias provenientes da Queimada, em sua maioria, se mudaram para a Rio Formoso.

No início da década de 80, uma pequena escola foi criada na aldeia para atender a crescente população que passava a ocupar a área. A Escola Indígena Municipal do Rio Formoso, então, foi responsável pela alfabetização de todos os habitantes atuais. Recentemente, há também Ensino de Jovens e Adultos.

Segundo levantamento sociolinguístico (Silva, 2013a), a língua indígena é aprendida por todos antes do Português, que é introduzido mais tarde na escola e quando a criança passa a visitar mais a cidade. Mudanças recentes, como a chegada da eletricidade em maio de 2009, apontam para o contato mais precoce da nova geração com a língua portuguesa, por intermédio da televisão e do rádio.

O contexto de uso do Português escrito está circunscrito ao contato com os não-índígenas, a atividades referentes aos órgãos de intervenção e apoio oficiais, como a FUNAI e a FUNASA, e a tarefas referentes à escola. O uso do Paresi escrito é legado à leitura de alguns poucos materiais de histórias existentes na língua, a bilhetes e, por vezes, a cartas entre parentes. Podemos dizer, então, que o uso do Paresi, verbal ou escrito, circunscreve-se apenas a atividades que não envolvem o não-indígena.

2.2. Documentação como um processo participativo: resultados

Ambos os projetos empreendidos tiveram uma abordagem participativa. Isso significa que os próprios indígenas participaram dos passos que envolvem o processo de documentação. Isso ocorreu em todas as esferas. Em reuniões, discutimos sobre o que era im-

portante ser documentado; em oficinas (na aldeia e no Museu do Índio) treinamos os interessados em técnicas de documentação (transcrição e tradução; registro em áudio e vídeo) e buscamos sempre estar atentos ao que os mesmos buscam com este projeto. Como resultado desse processo construímos o acervo que é resumido abaixo em (2):

(2)
100 horas de material audiovisual
12 horas de sessões de narrativas anotadas
35 horas de sessões de elicitação anotadas
1 gramática descritiva
Léxico com cerca de 2000 palavras
1 Centro de Documentação na Aldeia Formoso com equipamentos de ponta
2 pesquisadores indígenas treinados

2.3. Fluxo de trabalho: registro, metadata, transcrição e tradução, anotação e depósito

O trabalho em equipe é a força-motriz de um projeto de documentação cuja abordagem seja verdadeiramente participativa. Para isso, deve-se seguir um fluxo de trabalho bem definido. Primeiramente, devem ser tomados certos cuidados em relação ao **registro**. Além do que já foi mencionado e resumido na tabela 2, deve-se levar em conta a natureza dinâmica do evento a ser registrado. Por exemplo, uma festa tradicional é um evento cuja sequência de eventos, em geral, não é conhecida por membros externos à comunidade. Por isso, nesse caso, deve-se investir na autonomia e protagonismo da equipe de pesquisadores indígenas, que podem adiantar-se aos eventos e cobri-los de maneira mais natural e relevante.

O segundo passo da documentação é a posterior coleta de **metadata** (dados sobre dados) **básicos** imediatamente após a realização do registro. Essa medida é necessária para evitar a perda de informações que podem ser preciosas para definir o lugar e a relevância da sessão em sua base de dados. Além das informações essenciais do rótulo alfanumérico apresentado em (1), dados como participantes e uma descrição detalhada do evento são fundamentais. Essas informações, ainda que impressionísticas, serão fundamentais para a construção de uma base dados mais acessível para múltiplos usuários.

O terceiro passo é a **transcrição e tradução** do material coletado no ELAN. A transcrição de sessões de narrativas tradicionais deve ser feita por falantes nativos, visto que, como é o caso do Paresi-Haliti e de muitas outras nações indígenas, o conteúdo desses materiais usa vocabulário erudito, muitas vezes acessível apenas a uma pequena casta de iniciados. Isso garante traduções e glosagens mais precisas.

O quarto passo é a exportação das linhas de transcrição e tradução para uma base de dados de léxico e gramática. Atualmente há duas opções: Toolbox e FLEx, sendo a última plataforma a mais utilizada. Nessa fase, o linguista faz a **anotação** do material, segmentando-o e glosando-o. A partir desse trabalho, são gerados novos questionários para ses-

sões de elicitação.

O quinto passo é a **exportação** desse material transcrito, traduzido e anotado para o ELAN, onde vai ser juntado à informação audiovisual. Nesse estágio, a sessão está pronta para ser depositada.

O último passo é o **depósito** da sessão no acervo institucional. Nesse passo, decide-se, definitivamente, sua classificação na base de dados e discute-se seu grau de acesso juntamente à comunidade. Em geral, há três níveis básicos a serem escolhidos discricionariamente pela comunidade: completamente aberto, acessível mediante autorização de representante da comunidade e completamente fechado. A atualização e a inclusão de novos tipos de graus de acessibilidade pode ser feita a qualquer momento, conforme o desejo das comunidades envolvidas. Em (3) apresentamos um resumo do fluxo de trabalho apresentado:

(3) Fluxo de trabalho para cada sessão
registro > metadata > transcrição e tradução de sentenças>anotação>exportação>depósito

2.4. Treinamento de pesquisadores indígenas

Em um processo de documentação com abordagem participativa é importante que os membros da comunidade consigam "achar" o seu lugar dentre as várias atividades do projeto. Os principais campos de atuação são **o registro dos eventos** que envolvem habilidade para o manuseio de filmadoras, câmeras fotográficas e gravadores digitais, além de sensibilidade para a escolha de eventos que mereçam ser gravados e **o trabalho de transcrição, tradução e anotação** juntamente com o coordenador, que requer capacidade metalinguística e facilidade para desenvolver trabalhos com computador, já que isso é feito, na maioria das vezes, no programa ELAN.

Para a escolha de nossos consultores, tendo em vista esses dois campos de atuação, primeiramente, deixamos que os indígenas disponibilizassem-se, naturalmente, para um deles. Com a percepção inicial dessa inclinação, mostrávamos, informalmente, como funcionava o campo de atuação escolhido pelo aspirante a pesquisador. Os indivíduos que continuavam acompanhando curiosos o trabalho do coordenador eram convidados a participarem de uma das oficinas de capacitação, que eram, sempre, de caráter bastante prático.

As oficinas de áudio, vídeo e foto reuniam o grupo interessado nesse campo, que era instado a escolher registros relevantes de sua cultura. Assim, o grupo partia, imediatamente após as primeiras instruções, para a documentação de festas, histórias dos mais velhos e de outras práticas culturais relevantes. Em campo, com a orientação do coordenador, solucionavam dúvidas e resolviam problemas no próprio "fazer". As oficinas sobre o ELAN e o IMDI (gerenciador de *metadata*) necessitavam de um treinamento mais formal, pois a natureza desse aprendizado não é de caráter intuitivo. Os principais interessados foram professores que trabalhavam com a língua materna ou com história indígena na es-

cola da aldeia.

2.5. Estrutura (Servidor Museu do Índio) e *metadata* (*IMDI*)

A estrutura do acervo é a sugerida pela equipe do Museu do Índio em (4), baseada na concepção de Bruna Franchetto e Mara Santos para o projeto Dobes da língua Kuikuro. Sua concepção baseia-se na ideia de que o processo de documentação é contínuo, ou seja, não se encerra a um determinado projeto específico, e interdisciplinar, isto é, não abarca apenas a área de atuação de determinado projeto.

(4)
+---Arquivos
| +---Linguísticos
| | +---Elicitados
| | | +---Gravados
| | | | +---Estímulos
| | | | +---Lista de Palavras
| | | | +---Sentenças
| | | +---Não Gravados
| | | +---Lista de Palavras
| | | +---Notas de Campo
| | | +---Sentenças
| | +---Léxico
| | +---Uso Natural
| | +---Diálogo
| | | +---Gravados
| | | | +---Conversas
| | | | +---Entrevistas
| | | +---Não Gravados
| | | +---Conversas
| | | +---Entrevistas
| | +---Monólogo
| | +---Gravados
| | | +---Cantos
| | | +---Descrição
| | | +---Discursos Rituais
| | | +---Ensinamentos
| | | +---Explicação
| | | +---Narrativas
| | | | +---Históricas
| | | | +---Míticas

| | | | +---Pessoais
| | | +---Procedimentos
| | | +---Rezas
| | +---Não Gravados
| | +---Cartas
| | +---Descrição
| | +---Explicação
| | +---Narrativas Históricas
| | +---Narrativas Míticas
| | +---Narrativas Pessoais
| | +---Procedimentos
| +---Não Linguísticos
| +---Desenhos
| +---Imagens
| +---Músicas
| +---Instrumental
| +---Vocal
+---Estudos
+---Comparativo
+---Culturais
| +---Arqueológico
| +---Comparativo
| +---Etnográfico
| +---Geográfica
| +---Mapas
+---Linguístico
+---Afiliação Genética
+---Comparativo
+---Etnolinguístico
+---Gramática
+---Sociolinguísticos

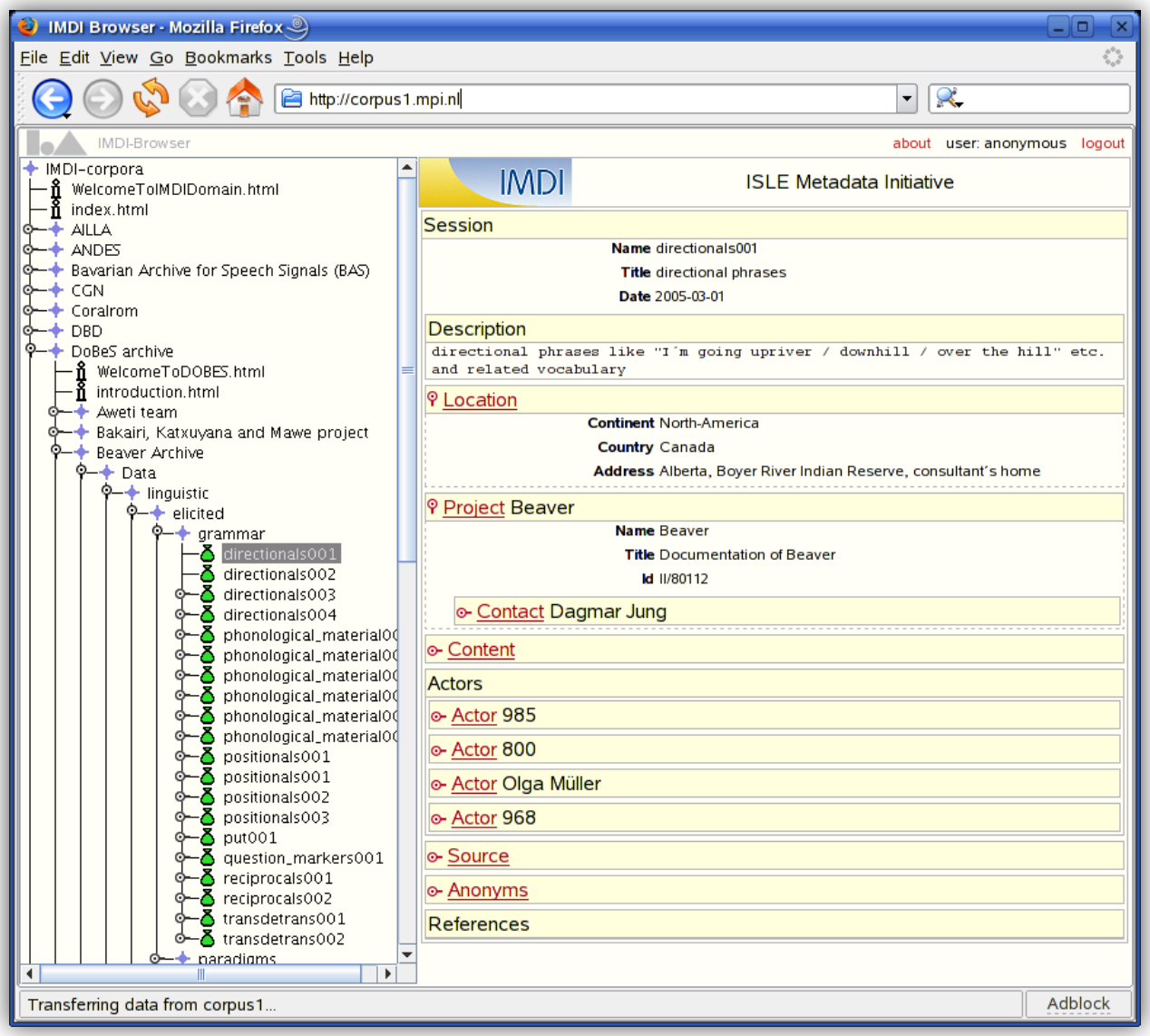
(Estrutura do acervo, PRODOCLIN – Museu do Índio, baseado em Franchetto & Santos para o projeto DOBES de documentação da língua Kuikuro)

Metadata são dados sobre dados. Eles permitem não só a identificação e recuperação rápida de sessões e de outros registros documentais, por meio de notações alfanuméricas convencionalizadas (cf. exemplo (1)), mas também a identificação de sua natureza, por intermédio de fichas. As fichas foram preenchidas com o editor de metadata IMDI[7]

[7] Atualmente, o IMDI já se encontra obsoleto (http://tla.mpi.nl/tools/tla-tools/imdi_browser**, Broeder, D., & Wittenburg, P. (2006)**. Um novo editor de metadata, Arbil, compatível com o IMDI, foi desenvolvido. http://www.lat-

(figura 3). Nele, além da identificação alfanumérica, foi inserido um nome descritivo para cada sessão. Essa medida tem como objetivo facilitar o acesso do público às sessões armazenadas no servidor (4), assim que as diretrizes de abertura terminarem de ser discutidas com a comunidade. Informações básicas que ajudam a identificar o conteúdo das sessões, assim como a que outros arquivos estão ligadas (transcrições, anotações, etc.) também são preenchidos.

Figura 1: IMDI Browser



O registro em áudio foi feito utilizando o gravador Marantz PMD-660 e o microfone Shure WH20 com entrada XLR e salvo no formato 44.1 Hz wav. Para os vídeos, utilizamos

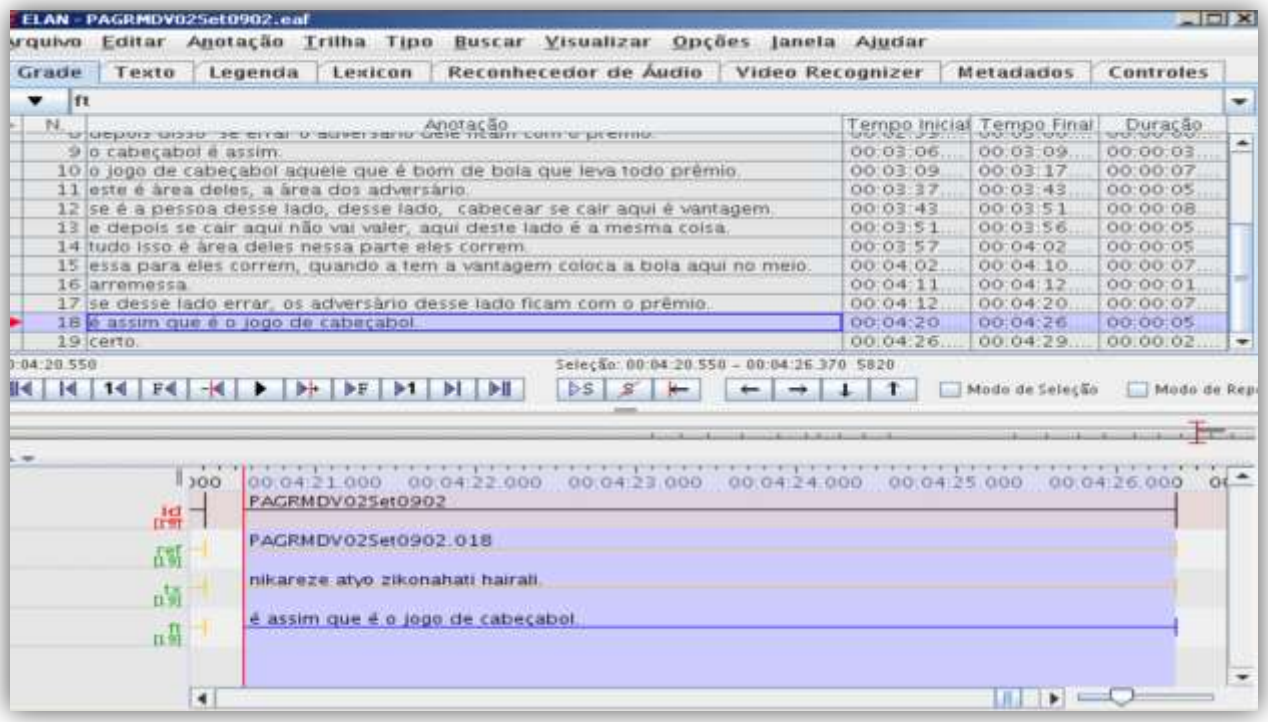
mpi.eu/tools/arbil/ pelo **Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. Withers, P. (2012).**

uma filmadora SONY com mídia mini-DV. Maiores informações técnicas podem ser encontradas facilmente nos sites do ELDP/SOAS e Dobes[8].

2.6. Textos transcritos, traduzidos e interlinearizados (*ELAN*) e base lexical (*Toolbox*)

Os textos foram transcritos e traduzidos em ELAN por um dos pesquisadores indígenas treinados (figura 2). Dessa forma, pudemos contar com traduções de um falante nativo. As linhas básicas que utilizamos no ELAN foram \id (sessão), \ref (linha de sequência), \tx (transcrição ortográfica) e \ft (tradução livre):

Figura 2: ELAN



Para a interlinearização (glosagem automática), utilizamos o programa Toobox[9] de gerenciamento de dados (figuras 3 e 4). Nesse programa mantemos três tipos de bases: uma de textos, uma de sentenças elicitadas e uma de léxico. Essa última (figura 4) alimenta a interlinearização das duas primeiras bases. Após a transcrição e tradução, os arquivos eram revisados e exportados para esse programa. As sentenças (\tx) foram glosadas (\gn), receberam segmentação morfológica (\mr) e informações sobre parte do discurso (\ps).

[8] http://dobes.mpi.nl/
[9] http://www.sil.org/computing/toolbox/downloads.htm

Figura 3: Base de textos Toolbox

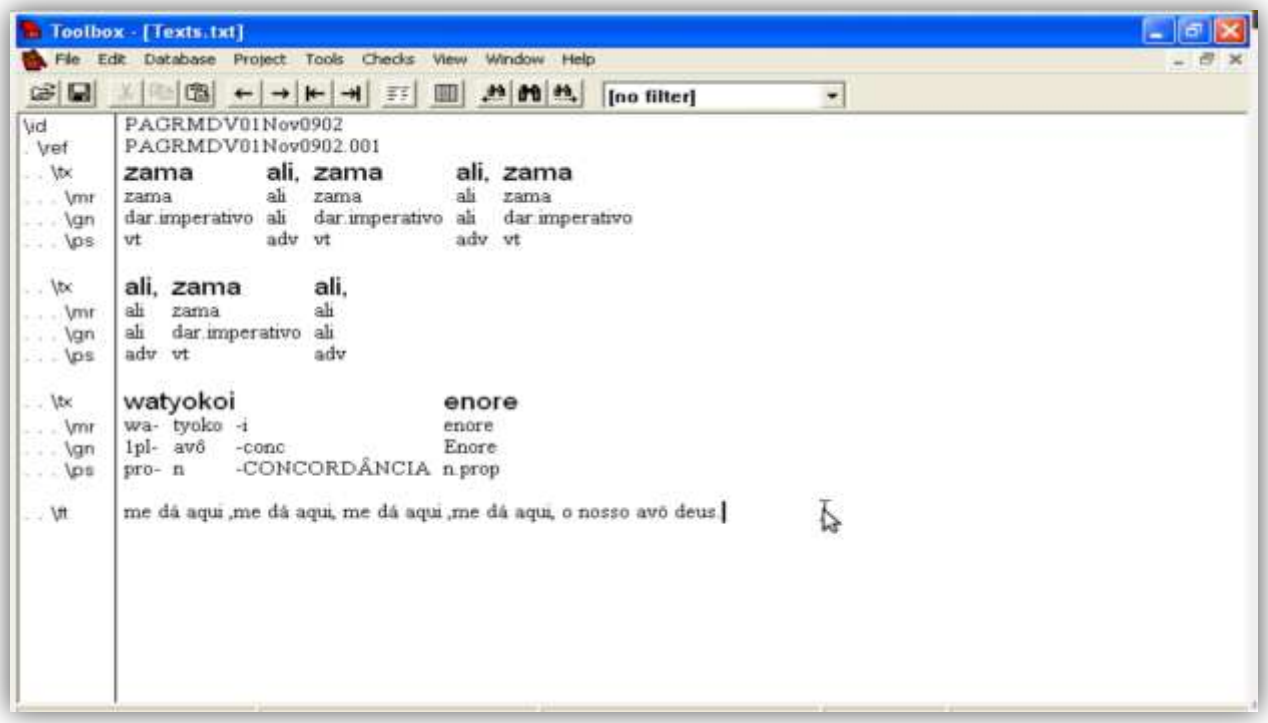
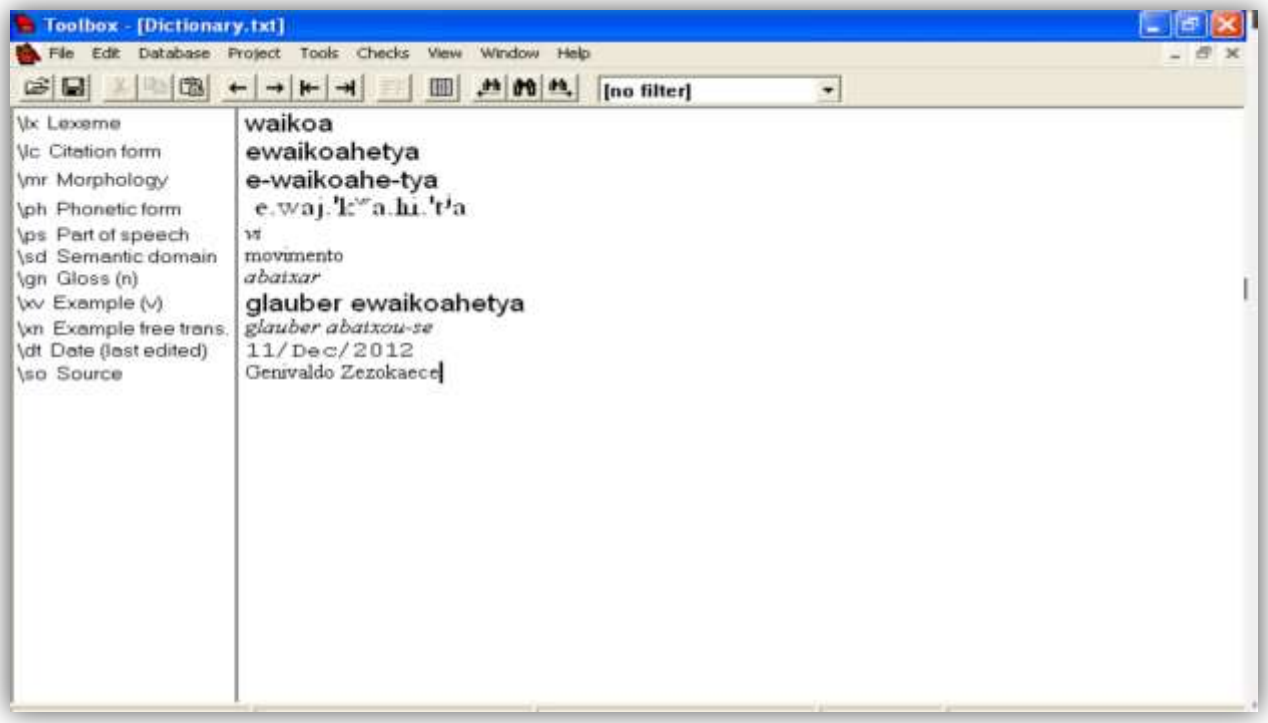


Figura 4: Base lexical Toolbox



2.7. Produtos gerados pelo acervo

Os principais produtos gerados pelo acervo foram uma gramática descritiva (Silva, 2013a), um dicionário piloto (Silva, 2013a), um livro de narrativas (Silva, 2014b) e algumas modificações na ortografia prática.

Redigimos uma gramática descritiva como um dos principais resultados de ambos os projetos. Esse trabalho aborda os principais aspectos de uma gramática de referência, sendo assim, foram exploradas a fonética e a fonologia (segmental e suprassegmental); a morfossintaxe das categorias lexicais maiores (nomes e verbos), categorias lexicais menores (adjetivos e advérbios), categorias funcionais do nível da sentença (negação, modo, aspecto); e a sintaxe de orações simples, coordenadas, relativas e subordinadas. É importante destacar que grande parte dos exemplos que ilustram a gramática foram retirados de contextos espontâneos, como narrativas.

Um produto como uma gramática descritiva representativa é um bom exemplo de atuação no que chamamos de **triplo impacto**. **Empiricamente**, por exemplo, revelamos alguns aspectos da diversidade tipológica do Paresi-Haliti, como o paralelismo morfológico revelado entre nomes, verbos e posposições, que apresentam o mesmo sufixo de concordância em paralelo à presença ou ausência de argumento externo (exemplos 5-11, Silva, 2014a).

Verbo inacusativo
(5a)
no=zan-**i**-∅
1sg=ir-**conc.1sg**-PERF
'eu fui'

(5b)
hi=zan-**e**-∅
2sg=ir-**conc**-PERF
'você foi'

Verbo transitivo
(6a)
no=tyoma-∅ (O)
1sg=fazer-PERF
'eu fiz (O)'

(6b)
hi=tyoma-∅ (O)
2sg=fazer-PERF
'você fez (O)'

Nome alienável
(7a)
no=kohatse<r>-**i**
1sg=peixe<CL>-**conc.1sg**
'meu peixe'

(7b)
hi=kohatse<r>-**a**
2sg=peixe<CL>-**conc**
'teu peixe'

Verbo inergativo
(8a)
no=tyoka-∅
1sg=sentar-PERF
'eu sentei'

(8b)
hi=tyoka-∅
2sg=sentar-PERF
'você sentou'

Posposição
(9a)
no=kako-**i**
1sg=com-**conc.1sg**
'comigo'

(9b)
hi=kako-**a**
2sg=com-**conc**
'contigo'

Nome inalienável
(10a)
no=kano
1sg=braço
'meu braço'

(10b)
hi=kano
2sg=braço
'teu braço'

(10c)
kano-ti
braço-N.POSS
'braço (avulso)'

Nome inerentemente possuído
(11a)
n=eze
1sg=pai
'meu pai'

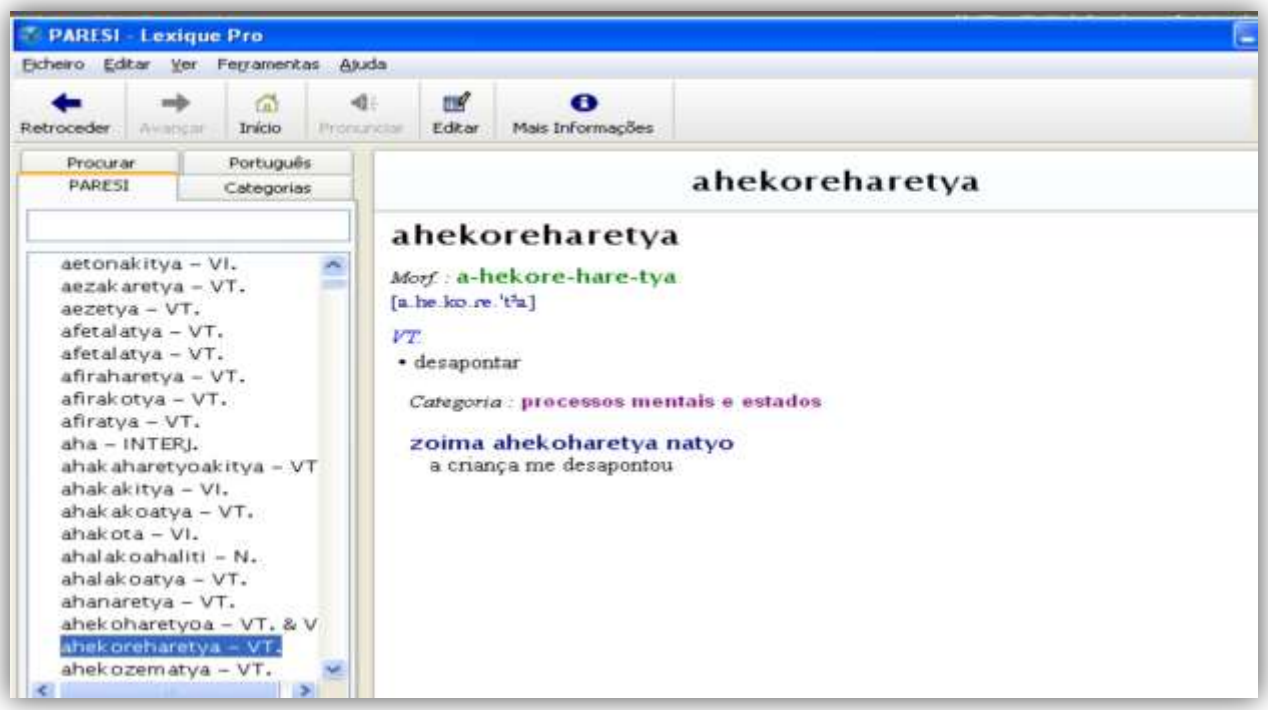
(11b)
h=eze
'teu pai'
(11c)
*eze-ti
pai-N.POSS
'pai'

Verbos inacusativos (5) e nomes alienáveis (7) e posposições (9) apresentam concordância; já verbos transitivos (6) e inergativos (8), nomes inalienáveis (10) e inerentemente possuídos (11), não; todas as posposições apresentam concordância, argumentamos que aquelas que não apresentam não o fazem por fatores independentes de caráter fonológico (cf. Silva, 2013a). O impacto **analítico** dessa generalização exocêntrica é capaz de falsificar a hipótese do Phase Impenetrability Constraint (PIC) (Chomsky 2000, p. 108), em que, somente o núcleo e sua fronteira ("edge"), Spec, são acessíveis a operações. Nesse sentido, somente predicados **sem** posição de especificador (verbos inacusativos e nomes alienáveis) ou que não constituam fase (posposições) pronunciam concordância. Já predicados **com** posição de especificador a bloqueiam (verbos transitivos, inergativos e

nomes alienáveis) (Silva, 2014a). Metodologicamente, o desafio não foi tão árduo, uma vez que praticamente todos os Paresi são bilíngues, o que torna o acesso a uma elicitação com fins descritivos gerais muito fácil. A questão metodológica, no entanto, é central em situações onde (i) não há língua franca entre o pesquisador e a comunidade e (2) o pesquisador tem objetivos estritamente teóricos.

O léxico Toolbox serviu para formatação de um dicionário piloto (Silva, 2013a, Apêndice 3), com entradas para segmentação morfológica, transcrição fonética, parte do discurso, glosa e exemplos. Para isso, utilizamos o gerador de dicionários *Lexique Pro* (figura 10).

Figura 5: Lexique Pro



A ortografia utilizada nas transcrições está em Silva (2009). Nela, propusemos mudanças em relação à ortografia de Rowan.

Tabela 3: Ortografia (Silva, 2009)

Fonema	Grafema	Exemplo
/b/	<b>	Aba “pai”
/t/	<t>	Tona “ele anda”
/tʲ/	<ty>	Atyakate “tronco de árvore”
/ɾ, d/	<r>	Tore “tucano”
/k/	<k>	Kahare “muito”
/m/	<m>	Ama “mãe”
/n/	<n>	Nanitya “eu como”
/h/	<h>	Haliti “gente”
/ɸ/	<f>	Ferakoati “dia”
/θ/	<z>	Zane “ele vai”
/j/	<y> em início de sílaba	Yakare “jacaré”
	<i> em final de sílaba	Kaiminiti “”
/w/	<w> em início de sílaba	Wahakanore “macaco”
	<o> em final de sílaba	Aolo “papagaio”
/l/	<l>	Alome “bugio”
/ts/	<ts>	Zoretse “estrela”
/a/	<a>	
/e/	<e>	
/i/	<i>	
/o/	<o>	

A nossa proposta ortográfica é técnica, pois buscamos conservar, ao máximo, a relação um-a-um entre fonema e grafema. Buscamos uma ortografia que preservasse a estrutura profunda em processos morfofonológicos. Em línguas com muitos processos morfofonológicos, como é o caso do Paresi, é mais vantajoso manter visíveis a forma subjacente de raízes e morfemas, pois isso auxilia a alfabetização mais sistemática e intuitiva de falantes nativos (Seifart, 2006). Um exemplo seria a grafia de (12a). O Paresi apresenta um processo morfofonológico de palatalização em que /tʲ/ é pronunciado como [ts] após [i] em fronteira de morfema. Para Rowan, a palavra seria escrita como em (12b). A nossa proposta contempla a forma (12c), que evidencia a fronteira de um morfema.

(12a)	(12b) Rowan	(12c) Silva
/na-ni -tʲa/ [na.nit.sa]	<nanitsa>	<nanitya>
1sg-comer-PERF		
'eu comi'		

O léxico, o dicionário piloto (Silva, 2013a) e o livro de narrativas bilíngue (Silva, 2014b) são exemplos de impactos preponderantemente **humanos**, uma vez que esse material pode auxiliar professores indígenas Pares-Haliti em suas aulas de língua materna com informações mais atualizadas em relação aos materiais anteriormente disponibilizados pela atuação de missionários na região (Rowan, O. 1961, 1963, 1964a, 1967b, 1972, 1975, 1977, 1978, 1983; Rowan, O. & Burgess, E., 1969).

3. A fronteira da era pós-90: o conceito de web 2.0

Retomamos nesta seção a pergunta feita no início deste trabalho, cujas possíveis respostas fundam o novo horizonte da área de documentação linguística: como a comunidade científica e as comunidades linguísticas envolvidas podem potencializar a **criação** de novos acervos anotados?

Os projetos Paresi-Haliti mostraram que é possível fazer muito em pouco tempo. No entanto, a concepção final de documentação é a mesma: gerar léxico, gramática e textos. O advento de novas tecnologias e a maior participação das comunidades indígenas no processo promoveram resultados mais robustos. No entanto, ainda é muito pouco se compararmos ao que temos para línguas majoritárias.

Em relação a novas concepções para a criação e atualização de *corpora* anotados, uma proposta (Silva, 2016, em preparação), baseado em Silva (2013b), pretende lançar mão de uma plataforma digital piloto com base no conceito de colaboração online em escala massiva da Web 2.0 para as comunidades que já têm acesso à internet. O mesmo conceito de autoconstrução e atualização é utilizado atualmente pela Wikipédia[10], que atualmente trabalha com o seguinte fluxo:

(13) o usuário consulta a Wikipédia> propõe edição para alguma lacuna apontada pela Wikipédia > os editores respondem > a alteração é validada pela comunidade usuária> vai para os fóruns> entra definitivamente no verbete.

A ideia é que os próprios membros da comunidade, ao utilizarem a base de dados do dicionário, por exemplo, sejam instados a preencher suas lacunas e a melhorá-lo, através de uma plataforma *online* amigável e que exija, preferencialmente, nenhum treinamento. Para isso, o linguista recebe essas alterações, as administra e as insere na plataforma. O fluxo de trabalho, no nosso caso, ocorreria da seguinte maneira:

(14) o pesquisador indígena utiliza a base lexical> o usuário editor (pesquisador indígena) propõe edição para alguma lacuna apontada pela base> a alteração é enviada para o linguista gestor> o linguista gestor avalia e insere na base> a informação fica disponível para os usuários> a nova informação é validada pela comunidade.

A interface da plataforma, ao sugerir alterações, deve fazer perguntas do tipo “esta tradução é boa? este exemplo é bom para essa entrada?”. A plataforma deve ser equipada com ferramentas que indiquem o *status* de qualidade da entrada (em revisão, definitiva, etc); as entradas lexicais, quando consultadas pela primeira vez por outro usuário, devem perguntar ao usuário editor “esta tradução realmente é boa?”, assim, a entrada passaria, por exemplo, de “em edição” para “definitiva” até que sua qualidade fosse contestada por

[10] https://pt.wikipedia.org/wiki/Wikipédia

outro usuário. Os usuários da base seriam identificados através de um *login* de sistema, dessa forma, toda a atividade individual seria registrada. Com o tempo, a fidedignidade individual de usuários poderia ser escalonada de acordo com alguns critérios, como número de contribuições e impacto das mesmas, como é feito atualmente pela Wikipédia.

A vantagem dessa plataforma é que isso possibilita a participação direta dos membros da comunidade, poder-se-ia alcançar um número muito mais rápido de entrada em um tempo muito pequeno; o processo de tradução de textos seria agilizado, com a participação de mais membros construiríamos um *corpus* mais consistente qualitativamente. A necessidade de elicitação de dados através de listas de palavras, algo maçante para os consultores indígenas, seria reduzida, além disso, a colaboração remota diminuiria bastante a necessidade de deslocamento na fase de processamento de dados.

Um exemplo concreto da utilização do conceito de colaboração massiva são os chamados códigos *captcha* (*Completely Automated Public Turing test to tell Computers and Humans Apart*)[11]. O código *captcha* consiste em solicitar que o usuário ao ser logado em algum *site* digite os caracteres de um arquivo de imagem. O objetivo expresso desse procedimento é assegurar que a tentativa de entrar no sistema está sendo feita realmente por um ser humano e não por um programa malicioso. No entanto, as imagens a que os usuários são instados a preencher manualmente fazem parte de bases de dados reais. Todas essas imagens são escaneamentos que não podem ser reconhecidos automaticamente por máquinas através de reconhecimento de caracteres (*Optical Character Recognition*, OCR) e que, portanto, devem ser preenchidos manualmente. A inserção de códigos *captcha* somam cerca de 150.000 horas de trabalho por dia, ou 75 anos de trabalho diário. Esse montante de trabalho "involuntário" está ajudando a digitalizar livros de maneira muito eficiente e revolucionária.

Projetos que se utilizaram desse conceito, como Duolingo[12], cujo foco era o aprendizado de línguas, provaram que, além da dinamicidade inerente a esse tipo de abordagem, que elimina o intermediário (ou diminui o seu papel), a qualidade do material produzido (i.e. traduções e julgamentos) são de maior qualidade. Assim, dar-se-ia um salto exponencial (ainda que não estritamente "massivo", dado o número de falantes e limitações técnicas das comunidades afetadas) não só na velocidade com que as entradas seriam adicionadas, mas também em sua qualidade e diversidade. Dessa forma, dicionários de línguas indígenas que, em geral, não passam de 2000-4000 entradas, poderiam conseguir 25000 entradas, o que é comum em versões, ainda que reduzidas, de dicionários de línguas familiares. O acesso a dados desse tipo certamente não só contribuiria para a pesquisa em linguística, mas também para a antropologia, a biologia, a arqueologia e as demais áreas.

Pelos motivos apresentados, isso tornaria a documentação mais eficaz e mais barata, com baixo custo operacional, vasta economia em viagens de campo, proporcionaria um salto quantitativo e qualitativo na construção de *corpora*, além de ser pedagogicamente

[11] www.captcha.net
[12] https://pt.duolingo.com/

interessante, altamente participativo e proporcionar a radicalização do conceito de autonomia, com a função do linguista sendo legada a de um mero gestor técnico. Uma abordagem como essa ajudaria a atender a dupla necessidade e ao triplo impacto supramencionados.

4. Conclusões: desafios de base

Nesta seção, pretendemos refletir sobre as seguintes questões nos possíveis desdobramentos futuros destes projetos de documentação: (i) como assegurar que as diretrizes éticas sobre o material culturalmente sensível sejam realmente respeitadas, tanto em suas dimensões técnicas, quanto políticas; (ii) como tornar emancipadora, de fato, a participação dos indígenas nesse processo, para além de tradutores e cinegrafistas treinados; e (iii) como tornar o processo de documentação mais eficiente e produtivo em termos técnicos.

A dimensão técnica de (i) diz respeito a como assegurar que o sistema de gestão de acesso não seja por demais burocrático e, ao mesmo tempo, mantenha a segurança. A dimensão política tem que fazer face a possíveis problemas como a apropriação unilateral dos meios de acesso por membros da comunidade (pesquisadores indígenas, lideranças, etc) quanto por (e principalmente) não-membros (coordenadores e demais acadêmicos). Compromissos que possam ir além de meios burocráticos que nos são legítimos, mas que podem ter outro valor representacional para outras formas de ver o mundo, são fundamentais de serem pensados. Longe de ser um relativismo oco, o que buscamos é evitar um *laissez-faire* negociado apenas com nossos princípios e forjado apenas para a legitimação dos mesmos. O impacto desse processo para as gerações futuras ainda nos é imponderável.

O que pode engendrar a consecução de (ii), em termos práticos, está em curso, cremos, em outras esferas. Recentemente, temos visto intelectuais indígenas ganhando formação acadêmica em universidades e cursos de pós-graduação. São antropólogos, linguistas e profissionais das mais diversas áreas, que assumem posições cada vez mais de destaque no ambiente acadêmico, como pesquisadores e professores. Caso a documentação permaneça como uma demanda para essas nações, é fundamental o fomento para que esses profissionais assumam a dianteira na coordenação e iniciativas nesses projetos.

Sobre (iii), devem-se pensar meios de tornar mais eficiente e barata a documentação, que se valem, muitas vezes, de recursos públicos, ao mesmo tempo que garantimos meios seguros de desenvolvimento para os itens (i-ii).

Cremos que observarmos esses três pilares, (i) ética, (ii) autonomia e (iii) eficiência, em nossas ações futuras é condição *sine que non* para que o esforço empreendido por todos os envolvidos nesse processo (indígenas e não-indígenas) não reproduza os mesmos vieses "assimilatórios" e "civilizatórios", que, de tão naturalizados, possam atuar inconscientemente na concepção e desenvolvimento dessas ações. Podemos errar pela ingenuidade de nossas concepções, mas nunca por negarmos a reflexão sincera e contínua sobre as mesmas.

4 Referências

Broeder, D., & Wittenburg, P. (2006). The IMDI metadata framework, its current application and future direction. International Journal of Metadata, Semantics and Ontologies, 1(2), 119-132. doi:10.1504/IJMSO.2006.011008.)

Davis, H., Gillon, C., & Matthewson, L. (2014). How to investigate linguistic diversity: Lessons from the Pacific Northwest. *Language*, *90*(4), e180-e226.

Gippert, J., NIKOLAUS, P. H., & ULRIKE, M. (2006). Essentials of language documentation (Trends in Linguistics. Studies and Monographs, 78.) Berlin: Mouton de Gruyter.

Gomes, A. Q. (2015). Línguas Indígenas Brasileiras: O novo campo de provas dos universais linguísticos. *LIAMES: Línguas Indígenas Americanas*,*15*(1), 149-165.

Hale, K., Krauss, M., Watahomigie, L. J., Yamamoto, A. Y., Craig, C., Jeanne, L. M., & England, N. C. (1992). Endangered languages. *Language*,*68*(1), 1-42.

Jakobson, R., & Boas, F. (1944). Franz Boas' approach to language. *International Journal of American Linguistics*, *10*(4), 188-195.

Krauss, M. (1992). The world's languages in crisis. *Language*, *68*(1), 4-10

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. Behavior Research Methods, Instruments, & Computers, 41(3), 841-849. doi:10.3758/BRM.41.3.591.

Lima, S. (2015). Resenha:'Trabalhando a partir de hipóteses falsificáveis' ou 'Sobre os mitos acerca dos C-linguistas': uma resenha crítica de "How to investigate linguistic diversity: lessons from the Pacific Northwest"(Henry Davis, Carrie Gillon e Lisa Matthewson). *Revista LinguiStica*, *10*(2).

Rowan, Orland. [sem data]. Formulário dos Vocabulários Padrões para Estudos Comparativos Preliminares nas Línguas Indígenas. SIL: Cuiabá.

___ (1961). A Phonemic Statement of Paresi. Cuiabá: SIL.

___ (1963). Parecis Discourse Structure. Cuiabá: SIL.

___ (1964a).High level phonology of Parecis – Preliminary version. SIL.

___ (1964b). Paresi Phonemes. Cuiabá: SIL.

___ (1967). Phonology of Paresi (Arawakan). Cuiabá: SIL.

___ (1972). Some Features of Paresi Discourse Structure. Cuiabá: SIL.

___ (1975). Wastudahenere Tahi: Histórias dos Nossos Estudos: Coleção de Histórias Escritas por Jovens. Cuiabá: SIL.

___ (1977). Estrutura Discursiva Parecis. Cuiabá: SIL.

___ (1978) (2001, edição digital). Iraiti Xawaiyekehalakatyakala: Dicionário Paresí-Português. Cuiabá: SIL.

___ (1983). Textos em Haliti (Parecis) I . Cuiabá: SIL.

Rowan, Orland; Burgess, Eunice. (1969) (2009, edição digital). Gramática Parecis. Cuiabá: SIL;

Silva, G. R. da (2009). *Fonologia da Língua Paresi-Haliti (Arawak)*. Dissertação de Mestrado, UFRJ, Programa de Pós-Graduação em Linguística, Rio de Janeiro.

___ (2013a). *Morfossintaxe da língua Paresí-Haliti (Arawak)*. Tese de Doutroado, UFRJ, Programa de Pós-Graduação em Linguística, Rio de Janeiro.
___ (2013b). Documentação de línguas: uma abordagem da web 2.0 (apresentação de trabalho) . In: Roessler, Eva- Maria & Coutinho-Silva, Thiago (orgs.). Atuais Tendências Potenciais na Documentação Linguística. Campinas – Unicamp , 2 a 5 de abril.
___(2014a). Nomes, verbos e posposições em Paresi-Haliti: uma generalização exocêntrica. VEREDAS online – Sintaxe das Línguas Brasileiras 2014/1.
___(2014b). Livro de narrativas Paresi-Haliti. Rio de Janeiro: Museu do Índio – FUNAI.
mmm
___(2016). How to build big corpora for linguistic documentation: a web 2.0 approach. (em preparação).
Seifart, F. (2006). Orthography development. In: *Essentials of language documentation*, 275-299.
Withers, P. (2012). Metadata management with Arbil. In: V. Arranz, D. Broeder, B. Gaiffe, M.
Gavrilidou, & M. Monachini (Eds.), Proceedings of the workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR at LREC 2012, Istanbul, May 22nd, 2012 (pp. 72-75). European Language Resources Association (ELRA).