Documentação Lingüística : O Formato de Anotação de Textos

Sebastian Drude[1]

[1]CCH – Museu Paraense Emílio Goeldi
Caixa Postal 399 – 66.040-170 – Belém – PA – Brasil
& Freie Universität Berlin – Alemanha
drude@museu-goeldi.br, sebadru@zedat.fu-berlin.de

Abstract. *This paper presents the methods of language documentation as applied in the Awetí Language Documentation Project, one of the projects in the Documentation of Endangered Languages Programme (DOBES). It describes the steps of how a large digital corpus of annotated multi-media data is built. Special attention is devoted to the format of annotation of linguistic data. The Advanced Glossing format is presented and justified.*

Keywords. *Language documentation; endangered languages; Advanced Glossing; linguistic annotation; multimedia corpus; Awetí.*

Resumo. *Este artigo apresenta a metodologia da documentação lingüística como aplicada no Projeto de Documentação da Língua Aweti, um dos projetos no Programa de Documentação de Línguas Ameaçadas (DOBES). O artigo descreve os passos de como um corpus digital abrangente de dados multimídia é construído. O foco especial é no formato de anotação de dados lingüísticos. O formato Advanced Glossing é apresentado e justificado.*

Palavras-chave. *Documentação lingüística; línguas ameaçadas; Advanced Glossing; anotação lingüística; corpus multimídia; Aweti.*

1. Introdução

No que segue, apresentarei os detalhes técnicos do fluxo de trabalho de documentação de textos no Projeto de Documentação da Língua Aweti (PDLA, ou simplesmente "Projeto Awetí"). Este projeto pertence à "lingüística documentacional", que vem se estabelecendo como um novo campo nestes anos. Por isso, na seção seguinte dou uma introdução a este novo ramo da lingüística.

O Projeto Awetí é um dos vários projetos do programa Documentação de Línguas Ameaçadas, "DOBES" (para o endereço eletrônico, vide referências abaixo), apoiado pela Volkswagenstiftung, Alemanha (independente da empresa do mesmo nome). Em geral, o fluxo de trabalho pode ser considerado como típico para projetos DOBES. Em alguns casos, no entanto, o Projeto Awetí segue suas próprias propostas, algumas das quais são parcialmente adotadas por outros projetos, especialmente por outros projetos sobre línguas do Alto Xingu (Trumai e Kuikuro; vide abaixo e os *links* no *site* do DOBES, sob DOBES nas referências).

O trabalho de campo no Projeto Awetí é desenvolvido pelo pesquisador principal, Sebastian Drude (ele iniciou a pesquisa entre os Awetí em 1998) juntamente com

sua colega Sabine Reiter (membro desde 2001, primeira estadia no campo em 2002). As questões mais gerais, incluindo questões metodológicas e teóricas, estão sob a responsabilidade do chefe do projeto, H-H. Lieb, e do pesquisador principal, Sebastian Drude.

O ponto principal da segunda parte deste artigo é a justificação, apresentação e aplicação do Advanced Glossing (AG), um formato geral de documentação de texto que foi desenvolvido por membros do projeto Awetí (Lieb e Drude 2000).

As idéias gerais do AG e sua implementação com o Programa **Shoebox/ Toolbox** já foram apresentadas anteriormente (vide referências, especialmente Drude 2002), aqui providencio somente um resumo. Porém, eu tratarei dos passos mais simples do fluxo de trabalho e dos detalhes dos argumentos que justificam o uso desse formato.

A partir da seção 3, este artigo é uma versão alterada e traduzida para o Português de um trabalho apresentado no segundo congresso do projeto EMELD, em Lansing, em 2003 (Drude 2003).

2. Lingüística Documentacional

A partir dos inícios da década de 1990, cada vez mais lingüistas defendem os seguintes pontos:

1. A diversidade lingüística que existe no mundo é uma riqueza inestimável, intimamente conectada com a diversidade cultural e conhecimentos tradicionais.
2. Esta diversidade lingüística está gravemente ameaçada e pode diminuir a 50% até 10% do atual dentro de poucas gerações (Krauss, 1992).
3. É urgente que se trabalhe contra esta perda, ajudando comunidades lingüísticas a manterem suas línguas tradicionais, e / ou documentando o máximo possível das línguas enquanto ativamente usadas ou lembradas.

Quanto ao primeiro ponto, já são geralmente bem conhecidas as cifras de que no mundo existem entre 5000 e 6000 línguas. É difícil dar estimativas mais exatas, especialmente porque é complicado decidir quais são os critérios exatos para diferenciar entre línguas e dialetos. Por exemplo, o "Ethnologue", o catálogo de línguas do Summer Institute of Linguistics, originalmente compilado para fins da missão evangélica fundamentalista, lista quase sete mil entradas na sua última edição (Gordon, 2005). Mas esta obra tem, em diversos casos, uma entrada para idiomas que usualmente são considerados dialetos de uma língua, e não línguas independentes.

As línguas no mundo pertencem a centenas de famílias lingüísticas de duas a várias dúzias de línguas (incluindo 'troncos lingüísticos', que abrangem várias subfamílias). Também há muitas línguas isoladas (sem línguas geneticamente relacionadas). A distribuição das línguas e das famílias lingüísticas é muito desigual. A grande maioria das línguas se concentra no assim-chamado terceiro mundo, especialmente nos trópicos da África, nas ilhas de Papua-Nova Guiné e da Indonésia, e na Amazônia. Há uma forte correlação entre a diversidade cultural e lingüística e a diversidade biológica, e vários fatos indicam que esta correlação não é um mero acaso (Nettle e Romaine, 2000).

Também varia muito o tamanho das comunidades que falam estas línguas. Mais do que a metade das línguas são faladas por menos que 10.000 pessoas, que juntos somam menos do que 0,3% da população mundial. Ao mesmo tempo, mais do que a metade da população mundial tem como língua materna uma das somente vinte línguas mais faladas (Gordon, 2005, Nettle e Romaine, 2000).

Nas Américas, atualmente há por volta de mil línguas. No Brasil, são ca. 170, sendo que aqui há na média entre 200 e 300 falantes por língua. Este número baixo, além de ser típico para povos nos trópicos (também em outras regiões do mundo), é também uma conseqüência da catástrofe demográfica nos Neo-trópicos depois de 1500, reduzindo a população de todos os povos. Somente a partir da segunda parte do século XX os povos sobreviventes conseguem se recuperar e aumentar sua população.

Muitos povos e línguas, porém, desapareceram nos últimos 500 anos. Há estimativas que no Brasil existiam, por volta de 1500, umas 600 línguas (Rodrigues 1993). Esta drástica redução exemplifica o segundo ponto, pois a pesar da recuperação demográfica, as línguas continuam desaparecendo, no mundo e no Brasil, e isto até num ritmo cada vez mais acelerado. Já agora há muitas línguas que não se passam mais para a próxima geração, e várias se usam somente em situações restritas, como em certas rituais.

A Perda da diversidade lingüística é universal (Dalby, 2003; Moseley, 2001). A causa principal é a exploração e integração de áreas remotas, na escala mundial (antigamente chamado "colonização", hoje: "globalização") e na escala nacional. Neste processo, as condições de vida dos grupos indígenas mudam-se drasticamente. Usualmente eles são marginalizados ou absorvidos. Há casos em que existem sanções abertas contra línguas minoritárias, variando de desvantagens econômicas e pressão sociais até a repressão declarada e chacinas.

Muitas vezes, o fator crucial é que a língua nativa é visto pelos próprios falantes como inferior e como impedimento na ascensão social. Supostamente para melhorar a situação dos filhos, a geração dos pais decide de não ensinar ativamente a língua tradicional à próxima geração. Um outro fator relacionado é a falta de escolaridade em línguas indígenas.

Desde os inícios dos anos noventa, o problema é percebido e discutido entre os lingüistas. A gravidade do risco é evidente. Não só é a perda nociva para a disciplina que perde seu próprio objeto de estudo e a possibilidade de conhecer os parâmetros e limites da variação das estruturas lingüísticas e de testar as teorias gerais sobre a linguagem humana. Em muitos casos o desaparecimento do idioma tradicional significa ao longo prazo uma perda dramática também para os próprios falantes e especialmente seus descendentes. Junto com idioma desaparece conhecimento tradicional de ordem ecológico e espiritual de inestimável valor que pode ser fundamental para a socialização e construção do indivíduo e a sobrevivência do grupo.

Por isso, programas para o fortalecimento das línguas nativas são de alta importância e urgência. Porém, a força dos fatores econômicos e sociais desfavoráveis bem como a falta geral de verbas nesta área são enormes, e o número de lingüistas e pedagogos com boa formação e interesse na questão é pequeno (Franchetto 2000; Moore, em prep.). Por isso, o enfraquecimento e desaparecimento de muitas línguas parecem irreversível e inevitável.

Algo que pode ser feito, no entanto, é documentar línguas enquanto ainda são usadas ou lembradas. A mesma época que vê o desaparecimento da diversidade também nos fornece com a tecnologia necessária para criar acervos abrangentes de línguas e culturas – hoje existe a possibilidade de fazer gravações digitais em áudio e vídeo de alta qualidade para custos acessíveis no 'campo', nos lugares onde as línguas e culturas são vividas no dia-a-dia.

Documentações neste novo sentido são diferentes da tradicional elaboração de gramáticas e dicionários, que descrevem a língua como sistema. As documentações da língua em uso permitem que hipóteses e análises sobre a estrutura lingüística sejam testadas e verificadas mesmo depois do desaparecimento dos falantes nativos. Elas também devem ser úteis para a comunidade dos falantes e seus descendentes, isto é, devem servir para atividades de fortalecimento ou revitalização da língua, devem apoiar iniciativas que visam à difusão de seu uso e ensino, inclusive em novas áreas como a comunicação escrita e a escola.

A partir da segunda metade dos anos 1990 surgem programas que visam à elaboração de documentações neste sentido. Alguns dos programas mais importantes são: DOBES (fundação Volkswagen, Alemanha), ELDP (SOAS, London), EMELD (pelos administradores da LinguistList), AILLA (U. Texas), PARADISEC (na Australia), LACITO (na França), LDA (pelo LDC, U. Pennsylvania), ELF (Yale U.).

Especialmente no início, estas iniciativas tiveram que estabelecer metas, métodos, tecnologias e padrões. Isto vale em particular para os oito projetos que participaram na fase piloto do programa DOBES. Entre estes, houve três projetos visando à documentação de línguas brasileiras – todas elas na área cultural do Alto Xingu: Kuikuro (Karib, por Bruna Franchetto, Museu Nacional / UFRJ), Trumai (língua isolada, por Raquel Guirardello, Instituto Max Planck Nijmegen & Museu Goeldi) e Aweti (Tupi, por Sebastian Drude, Freie Univ. Berlin & Museu Goeldi).

Hoje são mais do que 25 projetos em todos os continentes. Há agora também um projeto por Sérgio Meira (Leiden & Museu Goeldi) sobre três línguas brasileiras: Mawé (Tupí), Bakairí, Kashuyana (Karib). Desde o início em 2000 há um projeto tecnológico no Instituto Max Planck (MPI) em Nijmegen, Holanda. Este projeto desenvolve softwares e dá outro suporte para a documentação lingüística, como a digitalização e a preparação de arquivos de mídia (ver detalhes abaixo). É aqui também onde surge o acervo digital das línguas documentadas pelo programa.

Com o surgimento de uma quantidade considerável de projetos documentacionais e com o estabelecimento de normas e metodologias pode se falar em um novo ramo da lingüística, a Lingüística Documentacional (Himmelmann, 1998). O restante deste artigo apresenta alguns resultados e algumas propostas elaboradas no projeto de documentação da língua Awetí.

3. Visão geral da documentação lingüística

Uma documentação lingüística no sentido aqui usado é uma coleção de dados primários (gravações em áudio e / ou vídeo que contêm eventos de enunciação na língua, e possivelmente dados escritos) e dados secundários relacionados aos dados primários (informações sobre os dados primários, assim-chamados "metadados", e anotações e explicações do conteúdo). Usualmente uma documentação existe em forma

digital, isto é, em forma de arquivos num disco rígido ou em CD/DVD ou semelhantes. Este material é organizado em "sessões". Cada sessão abrange possivelmente diversos arquivos que contém dados primários e secundários, e obrigatoriamente um arquivo com metadados que especifica o conteúdo e os outros arquivos pertencentes à sessão. As sessões são organizadas / agrupadas conforme critérios significativos, p.ex. em forma de uma estrutura hierárquica.

Em geral, a criação de uma sessão que contem a documentação completa de um texto (dados da língua em questão com o máximo das anotações e metadados) no Projeto Awetí envolve os seguintes passos, que nem sempre necessariamente seguem uma ordem cronológica:

1. **a gravação em fita de eventos de fala;**
2. **a digitalização de gravações em arquivos eletrônicos (preliminarmente no campo, e depois pelo equipe do Instituto Max Planck em Nijmegen);**
3. **a transcrição ortográfica de textos na língua nativa;**
4. **a adição de tradução dos textos em Português, e depois em Inglês;**
5. **a adição de anotações de unidades morfológicas, sintáticas e do léxico;**
6. **a adição de anotações de estruturas morfológicas e sintáticas complexas, e também fonéticas e fonológicas;**
7. **a criação de metadados;**
8. **a conversão das anotações no formato digital final;**
9. **a inclusão de uma sessão na estrutura global do corpus.**

Há também questões de caráter geral cuja solução pode ser obtida a partir dos passos simples descritos acima. Estes incluem:

10. **as convenções para a organização e a nomenclatura dos vários arquivos conectados com a sessão, intermediários e finais.**

Todos esses passos, exceto passo (7), serão descritos nas seções seguintes. Como especificado acima, o foco da segunda parte do artigo será nos passos (5) e (6).

4. Dados primários: gravação e digitalização [passos (1) e (2)]

Desde 2002, todas as gravações de áudio e vídeo do Projeto Awetí foram feitas com dispositivos digitais (um walkman de MiniDisc (MD) Sony para gravação de áudio, e uma câmera mini-dv (MDV) Sony para vídeo, principalmente, com um microfone estéreo direcional Zennheiser MKE 300). O formato ATRAC interno ao mini-disco inclui compressão, o que em princípio significa uma perda de qualidade. Entretanto, há razões para acreditar-se que isso não afeta os métodos mais comuns de análise lingüística (cf. Wittenburg 2001). Agora há possibilidades de gravar digitalmente com dispositivos robustos a baixo custo, assim que não é mais recomendável o uso do MiniDisc comum.

Muitos dos dados considerados relevantes de um ponto de vista lingüístico (isto é, dados para serem anotados futuramente, conforme apresentado aqui) são baseados em gravações de áudio somente, especialmente diálogos naturais entre falantes nativos,

além de narrativas. Algumas sessões especiais planejadas foram gravadas em vídeo e áudio simultaneamente, o que criará problemas de sincronização dos diferentes sinais de transmissão. Por outro lado, a documentação de eventos culturais, que em geral não será mais anotada, é feita somente através de gravações de vídeo; estes evidentemente incluem um sinal de áudio que pode, no entanto, ser de qualidade inferior.

Gravações de áudio e vídeo são transformadas em arquivos eletrônicos (dados primários). Este processo é geralmente chamado "digitalização" – um termo algo equivocado já que atualmente a gravação é feita no formato digital e, depois convertida ou 'capturada' em um arquivo, de preferência, sem nenhuma conversão (intermediária) para um sinal analógico.

A criação final de arquivos de dados primários é feita pelo grupo TIDEL no departamento técnico do Instituto Max Planck em Nijmegen (vide link na página do DOBES, e em referências). O processo de digitalização é feito de acordo com as indicações (especialmente com relação aos pontos inicial e final das sessões simples) incluídas nos metadados criados para cada sessão.

Para fins da transcrição e o processamento adicional feitos ainda no campo, nós produzimos um arquivo de áudio preliminar das partes relevantes das gravações. No campo, nós usamos programas como o `Sound Forge` ou `Audacity` (vide referências) para criar um arquivo, geralmente sem qualquer edição além daquele de ajuste do volume do sinal. Como o walkman de MiniDisc permite somente a saída de um sinal analógico, não é de se admirar que o arquivo seja de qualidade inferior àquele criado exclusivamente através de transformações digitais.

Uma vez que a capacidade de armazenamento muitas vezes é restrita no campo, poder-se-ia pensar em usar formatos que envolvam compressão, tais como `MP3`, para o arquivo de áudio preliminar. No entanto, o `MP3` apresenta algumas distorções acústicas no início a cada vez que se toca uma gravação, e isso é particularmente perturbador, especialmente quando se ouve pequenas partes enquanto transcreve-se um texto. Por isso decidimos usar o formato de wave (Microsoft) não-comprimido (`.WAV`), e digitalizar somente as partes mais importantes das gravações de áudio. (Hoje, em 2005, há discos rígidos bem maiores e a possibilidade de gravar em DVD, assim que o espaço para armazenamento é cada vez menos restrito.)

As anotações (incluindo transcrição e tradução) são baseadas na segmentação dos arquivos de dados primários. Freqüentemente, os exatos pontos inicial e final neste arquivo não correspondem àqueles do arquivo final gerado no Instituto Max Planck (MPI), então uma discordância em fronteiras de segmentação é esperada. Preparei um script simples em emacs-lisp que permite fazer uma melhor sincronização das fronteiras de segmentação em arquivos de `Transcriber` e `Toolbox` (vide referências e abaixo).

5. Trabalho de campo com falantes nativos: anotação básica (transcrição e tradução) [passos (3) e (4)]

A fluência dos pesquisadores em Awetí, a língua estudada, é ainda incipiente. Portanto as transcrições e traduções dos textos (baseados em dados lingüísticos primários, conforme descrito acima) são feitas com a ajuda de falantes nativos. Nós sugerimos, no entanto, que em qualquer situação, até mesmo se os pesquisadores tenham um bom conhecimento da língua, que todas as anotações básicas sejam

checadas por ou com falantes nativos. Sendo assim, os procedimentos utilizados no processo de anotação básica são como segue.

Primeiramente usamos o programa **Transcriber** (vide referências) para dividir o sinal de áudio em segmentos que correspondam mais ou menos a sentenças. Até então, a segmentação é feita de forma impressionista, seguindo na maior parte das vezes critérios entoacionais. Isso poderá ser alterado quando a estrutura sintática da língua for mais bem entendida. Nosso objetivo é usar sentenças (completas ou elípticas) como unidades básicas para anotação de texto lingüístico, e não frases ou cláusulas, ou meros grupos de entoação. (Isso está de acordo com os princípios do Advanced Glossing; vide abaixo nas seções 9 e 10.)

No **Transcriber** ajustes de segmentação de tempo são relativamente fáceis de serem feitos, e incluem basicamente a divisão e fusão de segmentos adjacentes. Infelizmente, o mesmo não se aplica ao programa **Toolbox** (vide referências). No entanto, parece ser possível elaborar scripts pearl ou lisp para tornar o processo automático, mas isto ainda não foi utilizado no Projeto Awetí. Hoje em dia, o programa **ELAN** (vide MPI Tools) pode ser usado para a anotação com mais facilidade do que no início do projeto DOBES. Este programa permite o ajuste de tempo de várias anotações ao mesmo tempo. No projeto Aweti, não usamos o programa **ELAN** mais do que para fins experimentais.

O programa **Transcriber** é útil também para adicionar-se a transcrição diretamente a cada um dos segmentos (sentenças). No Projeto Awetí nós usamos a forma ortográfica, ao invés da fonética ou fonológica, para essa primeira transcrição. (Uma ortografia para Awetí foi estabelecida durante os últimos anos e está começando a ser usada por alguns falantes e na educação básica.) Além de ser muito mais rápido, esse procedimento evita dificuldades com a entrada de símbolos fonéticos (dependendo da configuração e o sistema operacional, **Transcriber** pode apresentar algumas dificuldades nessa área), e pode ser feito pelos próprios falantes.

Também viemos treinando alguns jovens Awetí para usar o computador e entrar transcrições com o **Transcriber**. De acordo com a nossa experiência, uma pessoa bem treinada pode transcrever pelo menos cerca de dois minutos em uma hora de trabalho, mas um minuto por hora é mais comum. Enquanto treinamos falantes nativos (isso inclui, em muitos casos, treinamento na aplicação da ortografia Awetí), nós calculamos menos de meio minuto por hora de trabalho.

Além de treinar falantes nativos, o processo de fazer transcrições oferece muitas oportunidades de discutir alguns pontos obscuros da ortografia, a qual pode ainda ser submetida a alguns esclarecimentos menores, adições e até mesmo mudanças.

Depois de transcrever o texto com a ajuda de ou por falantes nativos, a tradução é adicionada. **Transcriber** não foi projetado para fazer isso. No projeto Awetí, fazemos todo o processamento adicional usando o programa **Shoebox/Toolbox** (vide referências). **Toolbox** é uma versão mais atual do programa **Shoebox**, com algumas funcionalidades adicionais. Neste trabalho, nos referimos sempre ao **Toolbox**, mas em geral o exposto vale também para a versão anterior deste programa, **Shoebox**.

A conversão de **Transcriber** (este programa salva as transcrições em forma de arquivos XML) para **Toolbox** (este programa usa arquivos de texto pleno, organizado num 'formato padrão' do SIL), ou vice-versa, pode ser feita na maior parte

das vezes usando a ferramenta **Econv** desenvolvida pelo grupo TIDEL do Instituto Max Planck (MPI) em Nijmegen (vide "MPI Tools" nas referências).

Os bancos de dados **Toolbox** resultantes são de um tipo especial de banco de dados **Toolbox** (tipos de bancos de dados são praticamente um esquema "mark-up", isto é, eles são a descrição de tipos de campos permitidos em uma entrada, e o relacionamento de um com outro). Este tipo de banco de dados (**Econv.typ**) pode ser expandido para incluir linhas com diferentes anotações que se referem aos mesmos segmentos dos dados primários, principalmente linhas com a tradução aproximada em Português, a tradução palavra-por-palavra, e, mais tarde, a tradução em Inglês de cada sentença.

As traduções são obtidas de modos diferentes. No nosso caso, o modo mais comum é imprimir a transcrição com uma fonte grande e com muito espaço entre as linhas, e adicionar a tradução entre as linhas com falantes nativos (alguns em breve serão capazes de fazer isso por si mesmos). As traduções podem então ser digitadas no banco de dados **Toolbox** correspondente.

Uma outra forma é escrever a tradução enquanto a transcrição é feita. Isso atrasa um pouco o processo de transcrição, além de ficar difícil manter o controle do alinhamento de tempo. Se vários falantes estão envolvidos, isso pode ser um método válido, também para o treinamento destes em Português. Entretanto, especialmente quando o pesquisador trabalha com um só informante, e ao mesmo tempo faz a tradução (usando o **Transcriber**), uma boa solução é gravar a sessão de trabalho inteira e pedir ao falante nativo para traduzir cada sentença oralmente (isso é feito quase que automaticamente na maioria dos casos). Se alto-falantes são usados enquanto se transcreve (ao invés de fones de ouvido), alguém pode facilmente identificar o segmento na gravação e depois digitar a tradução diretamente no **Toolbox**, ouvindo a fita do processo da transcrição, possivelmente depois de retornar do campo.

Naturalmente, o processo de tradução dá acesso a uma variedade de informação, além de potencialmente levantar muitas questões e criar muitas oportunidades para a elicitação posterior, a qual deve ser feita o quanto antes possível, no campo.

O resultado desses passos é um texto de anotação mínima como foi acordado entre os projetos DOBES – isto é, uma anotação que inclui uma transcrição (aqui de caráter ortográfico) e uma tradução livre (pelo menos em Português, mas traduções em Inglês serão também providenciadas). Transcrição e Traduções têm referência temporal, isto é, há como identificar os segmentos nos arquivos de mídia aos quais se referem à transcrição e a tradução.

Os próximos dois passos, no. (5) e (6), são tratados mais detalhadamente na segunda parte deste artigo, na seção 10. Lá discutimos o caráter e o formato para anotações lingüísticas mais específicas, desde o nível fonético, via a morfologia até a sintaxe e semântica das sentenças do texto. Nesta parte, seguimos com os passos subseqüentes, que dizem respeito à conclusão das sessões e à criação do acervo como um todo.

6. Concluindo a sessão: conversão e inclusão dos dados no corpus [passos (7) a (9)]

Atualmente o projeto Awetí não usa o programa **ELAN** para anotação, uma ferramenta de anotação sendo desenvolvida no MPI em Nijmegen. Para nós os passos descritos acima, usando principalmente **Transcriber** e **Toolbox**, são suficientemente

eficientes. Não necessitamos de anotação direta de vídeo ou entrada de letras complexas UNICODE, funções que o ELAN oferece. As novas versões de Transcriber e o programa Toolbox, sucessor do Shoebox, também oferecem a possibilidade de usar o UNICODE, e a interação entre bancos de dados lexicais e dados de textos, o preenchimento semi-automático de anotação de tipos diferentes ('interlinearização', veja abaixo) são traços que nós precisamos continuar a usar e o ELAN precisará de algum tempo de desenvolvimento até ser capaz de substituir o Toolbox nesses aspectos.

No entanto, é possível converter os resultados de anotação de textos obtidos da forma descrita acima dentro do formato de dados ELAN, usando a mesma ferramenta Econv que também converte Transcriber para Toolbox, e vice-versa. Novas versões do ELAN importam dados do Toolbox. ELAN salva as anotações no formato XML que pode ser utilizado de outros programas. Os arquivos correspondentes têm a extensão *.eaf.

É obvio que para o propósito da apresentação, ELAN é fundamental, já que ele permite sincronizar imagens de vídeo e áudio simultaneamente com as linhas de anotação selecionadas. Nossa experiência com versões anteriores não tem nos dado resultados satisfatórios com relação à facilidade de operação para segmentar e entrar anotações, mas isso deveria ser basicamente resolvido com as novas versões.

ELAN ainda não suporta o formato Advanced Glossing (AG, ver seção 9) de uma forma ampla, mas futuramente permitirá o uso do formato (incluindo comentários, etc.) na sua totalidade. De fato quando começaram a desenvolver o ELAN, a proposta de AG foi utilizada como orientação básica para possíveis necessidades lingüísticas para uma ferramenta de documentação lingüística.

Entregamos os dados organizados em sessões para o arquivo do MPI, cada sessão sendo descrita por um arquivo de metadados (esta especificação de metadados segue o layout IMDI; vide IMDI 2003). Este arquivo IMDI contém descrições de detalhes técnicos das gravações e anotações, especifica seu conteúdo e identifica os arquivos relevantes que pertencem à sessão. Alguns dos arquivos de mídia são criados posteriormente de acordo com a informação dos metadados, a qual inclusive especifica os pontos iniciais e finais nos arquivos de áudio e vídeo dos arquivos finais a serem cortados dos arquivos mestres (originais) digitais. (Cada arquivo mestre – DMF, **D**igital **M**aster **F**ile – é uma cópia digital completa de uma fita inteira, enquanto os arquivos de mídia que pertencem às sessões são usualmente trechos destes arquivos mestres.)

Para criar o metadados usamos o Editor IMDI, também desenvolvido pela equipe do MPI. Embora a ferramenta se desenvolveu muito rápido como uma ferramenta de uso fácil, entrar os metadados continua sendo uma tarefa que consome muito tempo e que ocupa muito mais os nossos recursos de mão-de-obra do que planejamos no começo do projeto de documentação. No entanto, os dados serão procurados e acessados de acordo com critérios diferentes, e isso requer metadados completos.

O metadados, junto com os arquivos relevantes de mídia e anotação, são armazenados como parte do corpus, a parte central da documentação da língua; este está disponível via a Internet, dada a permissão de acesso das partes relevantes dos dados pela comunidade e informantes individuais. (Alguns dados terão de ter acesso restrito

ou até mesmo serão fechados para o público em geral.) Mecanismos que protegem direitos das partes envolvidas que ainda permitem o uso máximo dos arquivos estão sendo desenvolvidos, mas o tópico de direitos autorais, direito à imagem etc. continua sendo uma questão muito sutil, freqüentemente subestimada por projetos de documentação. A situação pode variar enormemente em diferentes países e continentes.

Os dados primários e secundários do Projeto Aweti do arquivo MPI serão organizados da seguinte forma:

1. **Material DE língua e cultura**
    a. **Dados lingüísticos**
        i. **Dados não-elicitados**
            A. **Dados monológicos (mitos, narrativas históricas, explicações culturais, textos de procedimentos, descrições etc.)**
            B. **Dados dialógicos (conversas, entrevistas etc.)**
        ii. **Léxico (inclui listas específicas de palavras)**
        iii. **Elicitações (frases e sentenças)**
    b. **Dados Não-lingüísticos (canções em outras línguas ou sem palavras, música instrumental, fotografias, desenhos feitos por falantes, material iconográfico etc.)**
2. **Material SOBRE língua e cultura**
    a. **Sobre a língua (sistema sonoro, ortografia, classes fechadas, esboço gramatical)**
    b. **Sobre o povo (informação etnográfica, informação sócio-cultural, informação histórica, informação geográfica (inclusive mapas), relações com outros grupos Xinguanos etc.)**

Se for necessário qualquer ramo final da estrutura acima pode ser subdividida. Isso serve especialmente para o ramo **1.a.i.A**, o qual inclui a maior parte dos dados lingüísticos. A estrutura de árvore acima apresentada é refletida por uma organização dos arquivos em diretórios e subdiretórios. Qualquer nó terminal contém quatro diretórios para os tipos diferentes de arquivos: arquivos de informações gerais, metadados das sessões, arquivos de anotação, e arquivos de mídia (áudio, vídeo e outros relacionados, tais como fotografias ou imagens e gráficos digitalizados).

Os metadados de cada sessão devem incluir a posição da sessão na estrutura de arvore do corpus. O corpus pode ser pesquisado, reorganizados e as sessões individuais podem ser exibidas na `WWW`, ou usando o `IMDI Browser`, mais uma ferramenta sendo desenvolvida pelo grupo TIDEL do MPI.

7. A organização de arquivos no computador [passo 10]

A organização de arquivos no computador é um aspecto geralmente negligenciado, ou é tomado como garantido. No entanto, lidamos com varias centenas de sessões, cada uma com arquivos de mídia (áudio e / ou vídeo etc.) e um ou vários arquivos de anotação, em formatos diferentes. Sem convenções e sem um esquema de organização, se perde facilmente nos milhares de arquivos.

Aqui oferecemos uma visão geral breve das soluções aplicadas no Projeto Aweti.

Há uma convenção de nomeação para arquivos mestres de dados primários (DMF) e as primeiras versões dos metadados (sessões) baseados nesses DMF, estabelecida no Consórcio DOBES durante sua fase piloto. Por exemplo, uma sessão baseada numa gravação de MiniDisc no Projeto Aweti tem o nome `AWSDAM23Jun0201-S01`, onde `AW` representa o Projeto **Aw**eti, `SD` o pesquisador **S**ebastian **D**rude (que fez a gravação), `AM` abrevia **A**udio-**M**iniDisc (`VDP` representaria uma gravação em **V**ídeo **D**igital no formato **PAL**), `23Jun02` a data da gravação. `01` indica que essa é a primeira fita (áudio ou vídeo) deste dia. Sendo assim, `AWSDAM23Jun0201` é uma etiqueta que designa uma fita de gravação. O `-S01` final identifica a primeira sessão contida ou baseada nesta gravação.

Vários arquivos, incluindo os arquivos digitais mestre e os arquivos de metadados são baseados nessa convenção e diferem principalmente na parte final `-S01` e na extensão do nome do arquivo. `AWSDAM23Jun0201-S01.IMDI` por exemplo é um arquivo de metadados que descreve essa sessão.

Outros nomes de arquivos são criados pelo próprio time Aweti, os nomes podem ou não incluir parte das convenções introduzidas acima, tais como data da gravação e numero da fita, mas geralmente algum elemento indicativo do conteúdo é adicionado para poder reconhecer o arquivo mais facilmente. Um exemplo é: `23Jun02-01-kawaka.sdb`, onde Kawaka é um nome (fictício) de um falante e `.sdb` é usado como extensão de arquivos de **B**ancos de **D**ados **S**hoebox (agora Toolbox).

A árvore dos arquivos nos discos rígidos de todos os computadores do projeto é a mesma para facilitar sua sincronização. Ela é organizada da seguinte forma:

`X:\Aweti\` contém cinco diretórios:

`Papers+Results\` (todos os resultados científicos, incluindo este artigo)
contém subdiretórios tais como: `2005-06-GEL`

`MPI-Corpus\` (a árvore de diretórios final, vide a última secção)

`Administration\` (orçamento, relatórios, autorizações, documentos oficiais...)
contém subdiretórios tais como: `VWS, FU, MuseuGoeldi, Internal, FUNAI`

`Media\` (arquivos de áudio, vídeo, fotografias etc.)

`Data\` (arquivos de anotação, metadata)

O conteúdo das pastas de `Media\` e `Data\` é organizado primeiro pelo ano, depois pelo tipo de arquivo. Por exemplo:

X:\Aweti\Media\2002\ contém, entre outros, os seguintes diretórios:

Audio-Field\ (arquivos de áudio intermediários, digitalizados no campo)
contém arquivos tais como: 23Jun02-01-kawaka.wav

Audio-DMF\ (Digital Master Files, arquivos mestres de áudio digitalizados no MPI, cada um corresponde a um MiniDisk)
contém arquivos tais como: AWSDAM23Jun0201.wav

Video-DMF\ (Digital Master Files arquivos mestres de vídeo digitalizados no MPI, cada um corresponde a uma fita Mini-DV)
contém arquivos tais como: AWSDAM23Jun0201.mpg

Audio-Sessions\ (arquivos de áudio cortados conforme o conteúdo)
contém arquivos tais como: kawaka-biogr.wav

Video-Sessions\ (arquivos de vídeo cortados conforme o conteúdo)
contém arquivos tais como: kawaka-biogr.mpg

Photos\ (fotografias ou gráficos e desenhos digitalizados)

Da mesma forma,

X:\Aweti\Data\2002\ contém, entre outros, os seguintes diretórios:

Transcriber\ (arquivos XML feitos com o Transcriber tais como:)
23Jun02-01-kawaka.trs

Shoebox-texts\ (anotações em formato de Toolbox, conforme descrito abaixo):
23Jun02-01-kawaka.txt, o arquivo Toolbox produzido pelo Econv
23Jun02-01-kawaka.sdb, anotações no formato AG-Syntax (vide abaixo)

Shoebox-lists\ (para listas de palavras transcritas ou elicitadas)

Print-out-versions\ (arquivos RTF para elicitação e tradução, tais como:)
23Jun02-01-kawaka.doc

Shoebox-lexical-databanks\ (para tabelas de glosagem morfológica, e para bancos de dados lexicais, tais como:) 23Jun02-01-kawaka-MGT.sdb

Metadata-work\ (arquivos de metadata que ainda precisam de revisão, tais como:)
AWSDAM23Jun0201-S01.imdi

Metadata-final\
(arquivos de metadata finalizadas, a serem incluídas na árvore final, tais como:)
kawaka-biogr.imdi

Além das pastas especificas para cada ano, em Data\ há ainda uma pasta General\ que inclui, entre outros, os três bancos de dados lexicais gerais do Toolbox.

Na pasta Administration\Internal há arquivos tais como tabelas Excel para o gerenciamento dos arquivos de média, das sessões, anotações, etc., para cada ano.

Com estes detalhes organizacionais fechamos a apresentação geral do passos do trabalho feito no Projeto Awetí. As próximas secções descrevem e justificam o formato usado para a anotação lingüística, Advanced Glossing, e sua implementação com o programa Toolbox.

8. Por que Traduções Morfêmicas Interlineares não são suficientes como um formato de anotação de texto

Parece ser unânime que a transcrição e a tradução (ao nível de sentença ou frase) sozinhas não são anotações suficientes para o propósito de documentação lingüística, especialmente no caso de línguas em perigo de extinção. Um formato de documentação completa deveria dar a informação necessária para desenvolver-se uma descrição mais profunda da língua, até mesmo se um dia já não houver nenhum falante nativo disponível. Isso significa que tem que ter a possibilidade de dar-se informação de diferentes tipos lingüísticos (pelo menos aqueles relacionados à estrutura da língua, tal como descrito numa gramática), cada um em seu próprio domínio.

Nossa proposta para um formato completo de anotação de texto é o Advanced Glossing (Lieb e Drude 2000, ver a próxima seção para detalhes). O termo *Glossing*, aqui traduzido ao Português por 'glosagem', se refere a anotações específicas a elementos isolados de um texto, como a palavras ou morfes. De fato, Advanced Glossing pode ser entendido como uma extensão elaborada de um formato de 'glosagem' bem conhecido, pelo menos em trabalhos que seguem os modelos funcionalistas ou tipológicos: Traduções Morfêmicas Interlineares (IMT do Inglês '*Interlinear Morphemic Translations*'), sistematizado primeiro por Christian Lehmann (1982).

Algum tipo de IMT é utilizado em muitos, se não na maioria dos projetos atuais de documentação lingüística. (Para uma visão geral sobre a variação de formatos indicados, vide Bow e outros (2003).) No entanto, eu argumentarei que esse formato apresenta problemas para este objetivo, embora já tenha sido provado que ele é bastante útil para fins ilustrativos, especialmente em gramáticas descritivas num modelo funcionalista / tipológico.

Traduções morfêmicas interlineares são, por definição, restritas à morfologia. Sua relação com a fonética, fonologia, sintaxe e semântica é indireta. Alguns autores perceberam esta restrição: como Bow et al. (2003) mostram, há expansões do formato que incluem glosas ao nível de palavra, além de substituir anotação ao nível de morfema. Ainda, na maioria dos casos a anotação é restrita a glosas que se relacionam a formas básicas isoladas – palavras simples ou morfemas simples.

Também no projeto EUROTYP, o formato foi alterado e expandido pra mais ou menos dez linhas com diferentes tipos de informação (Bakker et al., 1994). Mas há alguns problemas com este formato, os quais não podem, segundo nos parece, ser solucionados através da adição de mais linhas. Além da inclinação à morfologia (ou mais geralmente, à anotação relacionada à forma básica), estes problemas incluem a ausência de uma referência teórica em sua aplicabilidade a línguas de tipos diferentes, e a obscuridade na interpretação de glosas em casos de unidades gramaticais / funcionais.

Não existe tal coisa como 'descrição lingüística teoricamente neutra' ('*theory neutral*'), e o mesmo vale para a documentação, se ela é para ser mais do que simples gravação da fala. Qualquer anotação formula hipóteses, e qualquer hipótese é necessariamente formulada em termos de alguma teoria. Isso significa que também não pode haver nenhum formato de glosagem que seja neutro em relação à teoria. Porém, um formato eficaz deveria ser interteórico, isto é, deveria ser compatível com e ser capaz de ser usado por a maioria dos modelos de teorias lingüísticas.

De acordo com a classificação famosa de teorias lingüísticas formulada por Hockett (1958), há três modelos básicos de descrição lingüística: de um lado encontramos os modelos "Item and Arrangement" e "Item and Process", e de outro, o modelo "Word and Paradigm". O Estruturalismo Americano pertence ao primeiro tipo de modelo. O mesmo vale para a maioria das teorias que se têm desenvolvido a partir deste, incluindo abordagens tipológicas à descrição de língua (a maioria seguindo o modelo "Item and Arrangement"), e diversos ramos do pensamento teórico gerativista (muitos seguindo uma abordagem "Item and Process"). Essas teorias são tão difundidas que muitos lingüistas chegam até mesmo a ignorar a existência de outros modelos como, por exemplo, modelos do tipo "Word and Paradigm".

Este modelo não deve ser ignorado: durante séculos descrições lingüísticas, muitos delas melhor do que sua fama, foram baseadas em teorias que pertencem a este último tipo. Recentemente, há um interesse crescente de teóricos modernos em considerar a noção de paradigma como básico (isto vale em particular para algumas teorias neo-estruturalistas, especialmente na tradição européia do estruturalismo).

Um formato de glosagem geral deve ser compatível não somente com teorias baseadas nos modelos "Item and Arrangement" ou "Item and Process", mas também no modelo "Word and Paradigm". O formato IMT, no entanto, é claramente elaborado seguindo o modelo "Item and Arrangement". (Provavelmente, ele também é compatível com algumas variantes do modelo "Item and Process".)

Limitações do tipo de modelo subjacente são transferidas para os formatos de anotação baseados nele. Por exemplo, o modelo "Item and Arrangement" funciona bem para a maioria das línguas aglutinantes, mas há dificuldades com formas sintéticas de palavras em línguas flexionais, e com formas analíticas em línguas de todos os tipos tipológicos. Essas dificuldades são a principal razão pela qual não podemos recomendar IMT como o principal instrumento para dar informação gramatical em anotações a textos em documentações lingüísticas. Para discutir o problema, vejamos alguns exemplos. Como Lehmann (1983), eu usarei Latim como exemplo, e até mesmo variações de alguns dos seus exemplos.

O exemplo abaixo é típico de IMT, com glosas em Inglês:

(1) ***time*** ***-o*** ***ne*** ***veni*** ***-a*** ***-t***
fear -1.SG NEG.VOL come -SBJV.PRES -3.SG
'I am afraid he might come.'

Para uma documentação lingüística, essa anotação pode ser considerada incompleta, uma vez que as gramáticas do Latim descrevem paradigmas verbais que envolvem diversas categorias para qualquer forma verbal finita, no mínimo pessoa, número (verbal), genus verbi (*voice*), modo e tempo. Na sentença em (1) temos, por exemplo, a ocorrência da forma ***timeo***, uma forma que, conforme as descrições usuais, pertence às seguintes categorias: Primeira Pessoa, Singular (verbal), Indicativo, Ativo, Tempo Presente. (Outras terminologias falam aqui não de categorias, mas de 'traços morfossintáticos', '*morphosyntactic features*', um termo que é freqüentemente aplicado, mas que por sua vez não tem, no meu conhecimento, sido claramente definido em lugar algum.) Algumas dessas categorias podem ser razoavelmente relacionadas à ocorrência do afixo ***-o***. (Desse modo, ***-o*** seria analisado como um morfema portmanteau de valor

fortemente funcional.) Para representar esses fatos, a glosagem poderia ser aumentado da seguinte forma:

(2) ***time -o***
fear -1.SG.IND.AC*T*

A questão é: onde indicamos que ***timeo*** é uma forma do Tempo Presente? Não podemos assinalar isso à ocorrência do afixo ***-o*** já que o mesmo ocorre em formas de outros tempos; por exemplo, ***time-b-o*** (Futuro 1), ou ***tim-u-er-o*** (Futuro 2).

Lehmann oferece duas possibilidades para tais casos, as quais são para serem entendidas como variantes de anotação de uma análise que inclui a ocorrência de um morfema zero:

(3) ***time -Ø -o*** ***time -o***
fear -PRES -1.SG.IND.ACT fear(PRES) -1.SG.IND.ACT

Como no caso de morfemas portmanteau, essa análise pode ser discutida, mas certamente ela não é neutra teoricamente, já que muitas teorias rejeitam morfemas 'zero' (descrições tradicionais de Latim assinalariam a forma do tempo presente à ocorrência de *time*, uma forma do tema específica para o tempo Presente). Mesmo que algumas interpretações válidas das glosas em termos de outras teorias (as que, por exemplo, não postulam morfemas 'zero') sejam imagináveis, ainda temos que concluir que há problemas do formato IMT com relação à neutralidade teórica, ou melhor, interteoreticidade, pelo menos com relação a formas sintéticas de palavras em línguas flexionais.

O mesmo se aplica, ainda mais fortemente, ao caso de formas analíticas de palavras. Considere a seguinte sentença:

(4) ***monitus eram ut venirem***
'I had been asked to come.'

Aqui temos, novamente de acordo com descrições tradicionais, a ocorrência da forma verbal analítica (ou perifrástica) ***monitus eram***. Sete categorias poderiam ser atribuídas a esta forma: Primeira Pessoa, Singular (verbal), Indicativo, Passivo, Mais-que-Perfeito, Nominativo (verbal), Masculino (verbal). As últimas duas podem ser questionadas, mas isto não é o nosso ponto aqui. Então, como seria uma Glosagem Morfêmica Interlinear que reflita esta análise?

Considere a seguinte possibilidade de glosagem para a sentença em (4).

(5) ***moni -t -us*** ***er -a -m***
ask -PART.PF.PASS -NOM.SG.M PASS -IND.PAST -1.SG.ACT

ut veni -re -m
[??] come -SBJV.PAST -1.SG.ACT

Algumas das categorias a que ***monitus eram*** pertence aparecem diretamente sob uma ocorrência de um morfema específico (cf. Primeira Pessoa, Indicativo, Nominativo, Masculino). Outras aparecem mais do que uma vez (Singular, Passivo; a redundância está no objeto, não na descrição). No entanto, numa teoria "Word-and-Paradigm", todas as categorias se aplicam à forma (analítica) inteira, e não somente a um morfema. Porém, não há nenhuma maneira fácil de reconhecer que as duas palavras formam uma

forma analítica (e que esta forma, abrangendo as duas palavras, pertence à mesma categoria juntas). Há propostas que sugerem o uso de colchetes para esses casos. Mas o que fazer no caso de ocorrências descontínuas? Considere uma possível variante da sentença (4) acima:

(6) ***monitus ut venirem eram***

Aqui, colchetes se tornam impossíveis a não ser que se utilizem recursos especiais (tais como indexação), os quais resultam em uma anotação de difícil recepção, sem mencionar as dificuldades para uma implementação digital.

Mesmo que essas dificuldades fossem de alguma forma superadas, ainda temos um problema similar aos problemas com ***timeo*** acima: onde indicaremos o Tempo Mais-que-Perfeito? A forma toda pertence a esta categoria devido à sua composição: a combinação de uma forma do Particípio Perfeito com uma forma do verbo auxiliar no Tempo Passado. Cada uma destas propriedades é indicada nas glosas, mas não há como assinalar a categoria Mais-que-Perfeito à forma analítica do verbo como um todo: o formato IMT é restrito a glosas de morfemas (ou de palavras simples). Em Latim, porém, Mais-que-Perfeito é uma categoria gramatical que contém formas analíticas. É por isso que essas categorias (os 'traços morfossintáticos') de fato não são morfológicas, mas sim sintáticas.

Há outros problemas com o formato IMT, for exemplo em conexão com os assim-chamados 'morfemas livres' (por exemplo, eu não forneci, de propósito, qualquer proposta de glosa para ***ut*** acima). Outro ponto crítico é que o caráter (morfológico ou sintático, categorial, relacional ou semântico) para a glosa é freqüentemente obscuro. A lista poderia ser continuada.

Novamente, essas são questões cruciais que qualquer teoria tem de enfrentar em seus próprios domínios; o ponto é que o formato de glosagem IMT não somente é longe de ser neutro teoricamente (nenhum formato de glosagem o é), mas também não é suficientemente interteórico, porque é simplesmente incompatível com muitas teorias.

Um passo importante em direção a essa relação interteórica é fazer com que todas as suposições tácitas e conceitos e convenções dependentes da teoria sejam o mais explícitas possíveis, de forma que seguidores de outras teorias possam interpretar a anotação em seus próprios termos. Entretanto, infelizmente as tentativas de formalizar a IMT falham nesse respeito também. (Isso pode ser uma conseqüência das teorias nas quais ela é baseada.) Quando se quer saber o que exatamente as glosas em letras capitais significam, as respostas são diversas e muitas vezes só intuitivamente compreensíveis.

Consultando Lehmann (1982) ou as instruções de EUROTYP (Bakker et al. 1994), essas glosas são explicadas como "rótulos de categoria gramatical". Mas elas representam categorias? E se sim, "gramatical" significa "sintático" ou "morfológico"? Encontramos também nos mesmos textos explicações que as glosas expressam o "significado" de "elementos" gramaticais (i.e. morfes, ou morfemas, ou de que?), ou indicam a "função gramatical" desses elementos. Afinal de contas, eles representam informação categorial, semântica, funcional, ou de que tipo?

Lamentavelmente, em muitas descrições individuais de línguas, a única explicação para esses elementos ('traços morfossintáticos') é a versão longa para as

abreviações. (Isso pode ser apropriado desde que o foco seja nas próprias abreviações, cf. Croft 2003 que evita qualquer afirmação ontológica.)

Uma inspeção geral da lista de abreviação do projeto EUROTYP (vide Lieb et al., 2001) demonstrou que os ca. 550 rótulos referem-se a entidades lingüísticas que são geralmente interpretadas como pertencendo a tipos ontológicos muito diferentes: (i) quase a metade refere-se a categorias "morfológicas" (ou, conforme argumentamos acima, sintáticas, pelos menos aquelas relacionados ao paradigma de palavras); (ii) mais ou menos uma centena refere-se a classes lexicais de palavras, (iii) ca. 60 são termos para relações sintáticas; (iv) outras se referem a categorias de constituinte sintático; (v) outros a papéis semânticos; e (vi) outros designam propriedades de ordem de palavra, variedades, tipos de sentença, ou outras. Pior, cerca de 75 parecem complemente obscuros ou não específicos. Não se pode excluir a possibilidade que isso no futuro atrase o uso de textos anotados em EUROTYP.

O ponto da discussão aqui é que essa imprecisão tem que ser evitada em documentações de línguas, para que futuros lingüistas possam fazer uso das documentações, possivelmente aplicando teorias que nós hoje nem podemos imaginar.

A proposta de uma "ontologia" (um banco de dados de conhecimento geral) tal como GOLD (vide Farrar and Langendoen 2003) pode ser um avanço prometedor de evitar ambigüidade terminológica, mas corre o risco de fornecer somente mais uma teoria lingüística (afinal, por uma grande parte, teorias consistem em definições de termos), ou de ser um amalgamo heterogêneo de teorias atualmente mais populares ou teorias restritas a uma comunidade particular. A compatibilidade da ontologia proposta com teorias Word-and-Paradigm teria de ser cuidadosamente checada. De qualquer forma, precisamos de explicações explícitas de termos, e a ontologia GOLD pode servir como um ponto de referência para qualquer teoria particular. (É num sentido semelhante que o Advanced Glossing é uma proposta para um ponto de referência para formatos de glosas.)

9. Traços básicos do Advanced Glossing e sua aplicação no Projeto Awetí

Em Lieb e Drude (2000, disponível na *Internet*) fazemos uma proposta para um formato de anotação de texto chamado Advanced Glossing. Advanced Glossing (AG) foi elaborado para a documentação lingüística e é compatível com a maioria, possivelmente com todas as teorias lingüísticas. AG não é um modelo de dados, mas sim um formato da forma de apresentação; em princípio, ele é compatível com diferentes modelos gerais de estrutura de dados para textos interlineares, tais como aquele proposto por Bow e outros (2003).

AG nunca foi projetado para ser um esquema obrigatório, do tipo que qualquer projeto de documentação lingüística deveria seguir; mas AG pode servir como um modelo máximo de referência, providenciando um ponto comum para a comparação de diferentes esquemas de anotação. Sendo um modelo máximo, qualquer tipo de informação lingüística que se possa imaginar tem seu devido lugar. Uma anotação concreta de um texto pode usar um subgrupo das "linhas" que propomos. Quem quiser aplicar o esquema completo, pode fazê-lo aos poucos, deixando claro onde falta informação por falta de conhecimento.

As idéias básicas do AG incluem a separação estrita da informação sintática da morfológica. Há glosagens para cada sentença de um lado, e para cada palavra gramatical do outro, cada uma em forma de uma tabela (*glossing table*). Tabelas de glosas sintáticas e morfológicas são análogas em muitos aspectos, mas a informação fonética (dos segmentos, sua estrutura e a entonação) é dada somente em tabelas de glosas sintáticas (evidentemente, estas podem tratar de um enunciado de só uma palavra, por exemplo no caso de elicitações de palavras isoladas). Glosagens sintáticas não incluem qualquer informação sobre a composição morfológica interna das palavras individuais. Esta é fornecida nas tabelas correspondentes de glosagens morfológicas.

Cada tabela de glosas consiste de 13 linhas (numa versão revisada que está em preparação, duas ou três linhas serão adicionadas). Algumas linhas são **holísticas** (tais como uma tradução livre de uma sentença), ou podem consistir em listas (como as linhas para estruturas de constituintes que permitem lidar com constituintes descontínuos). Várias linhas, no entanto, consistem de **células** que correspondem a uma unidade básica (*base form*) individual. As unidades básicas no caso de glosas sintáticas são as palavras, no caso morfológico são os morfes individuais. Linhas com células incluem aquelas para informação semântica de cada unidade básica. Esta informação semântica é separada da informação sobre categorias. Diferenciamos entre dois tipos de categorias: categorias lexicais (tais como classes de palavras; por exemplo, categorias 'POS'), e categorias de formas (tais como casos, tempos, etc., no caso sintático). Esses tipos de categorias são colocados em duas linhas diferentes. Cada destas linhas consiste de células, sendo que cada célula pode conter uma lista de categorias.

Para cada tabela, linha ou célula pode-se acrescentar um comentário expressando dúvida ou dando alguma explicação. Células ou linhas podem ficar vazias, seja por escolha ou porque não há ainda a informação relevante. Para outros detalhes referimos o leitor à proposta original em Lieb e Drude (2000). Uma segunda versão desse formato está em preparação e incluirá umas poucas extensões, além de fornecer uma explicação mais detalhada sobre o formato. A maior parte da discussão do formato IMT na última seção será incluída na versão revisada.

No Projeto Aweti, aplicamos o AG como quadro de referência em qualquer anotação de texto. Isso não significa que em um texto haverá glosagens sintáticas completas para todas as sentenças e uma glosagem morfológica para cada palavra. Ao contrário, glosagens completas são planejadas somente para poucos minutos de texto gravado, pois o processo de preencher todo os tipos de informação consome muito tempo.

Para um sub-corpus menor (possivelmente poucos horas de texto), planejamos providenciar tabelas básicas de glosagens sintáticas e morfológicas, sendo que cada tabela irá conter cerca de 8 linhas, parcialmente ou completamente preenchidas, além de algumas linhas adicionais derivadas (tais como a tradução em português e a tradução em inglês) e comentários.

A maior parte do corpus consistirá de sessões lingüisticamente relevantes com anotação mínima (isto é, somente transcrição e tradução), ou de dados relevantes culturalmente (sem transcrições) com somente alguns comentários. Num projeto de documentação de cinco anos, com pouca mão-de-obra disponível (incluindo falantes nativos que estão sendo treinados durante o projeto), não esperamos poder oferecer anotação básica para a maioria do material rico de narrativas orais (principalmente mitos) dos

Awetí, num total de 70 a 100 horas, 60 das quais esperamos poder pelo menos gravar e digitar.

Como foi dito acima utilizamos o programa **Toolbox** para anotar textos e reunir a informação sobre unidades lexicais (palavras e afixos). Em Drude (2002) descrevi uma configuração relativamente complexa do **Toolbox** usada para implementar o AG, e eu refiro o leitor a essa descrição. Cada texto anotado corresponde a um banco de dados do **Toolbox**, onde cada entrada contém uma tabela de glosagem sintática, juntamente com os comentários; bancos de dados de um segundo tipo contêm todas as glosas morfológicas. As propriedades dos campos do **Toolbox** são configuradas de tal forma que refletem o caráter ontológico de cada linha de glosagem. O mecanismo de interlinearização do programa é usado para implementar o alinhamento em colunas e daí a estrutura de células de certas linhas. Além disso, há a possibilidade da interação com bancos de dados lexicais (para os quais o **Shoebox** foi originalmente elaborado). Desse modo o preenchimento semi-automático de certas informações (significados lexicais, informações de categorias, etc.) é possível: para preencher células de tabelas de glosa sintática, a informação é recuperada das glosas morfológicas relevantes, e para estas os processos de busca utilizam bancos de dados lexicais. Alguém pode também pular para a relevante entrada lexical diretamente das glosas sintáticas.

Depois destas informações gerais sobre AG, podemos providenciar a descrição dos passos concretos no fluxo de trabalho no Projeto Aweti. Esta descrição possivelmente fará os últimos parágrafos algo mais concretos.

10. Adicionando anotação lingüística [passos (5) e (6)]

A anotação básica como descrito acima (vide seção 4) resulta num banco de dados **Toolbox** com entradas para as sentenças individuais, sendo que cada entrada inclui uma transcrição ortográfica, a tradução em português (e em inglês) e uma marcação temporal, isto é, uma referência única a um ponto inicial de um segmento correspondente num arquivo de áudio subjacente.

O banco de dados é ainda do tipo **econv.typ**, o tipo de banco de dados que vem com o programa **econv**, com definições adicionais para diferentes campos de tradução. O próximo passo é então a conversão desses bancos de dados para o tipo criado para glosagens sintáticas no formato AG. De fato, a conversão ao formato **AG-Syntax.typ** é obrigatório para os processos de exportação mencionados anteriormente, inclusive para imprimir as transcrições.

Escrevi mais um script simples em **emacs-lisp** para automatizar o processo de conversão, incluindo a conversão para marcar acentos nasais, e a conversão dos marcadores de campo. (Todos os scripts mencionados neste artigo podem ser obtidos com o autor.) O resultado é exemplificado pela seguinte entrada de dados:

```
\ref      0002.690
\per      mawalaja
\SXII     jatätsu jatä ozoporywyt;
\SXIIIn1  assim é que é nossa tradição:
\SXIIIe   it is like this that our tradition is:
\dt       02/May/2002
```

Como dito acima, para a maior parte dos dados isto será toda a anotação que é fornecida. Cada linha representa uma linha de uma tabela de glosagem sintática. As

transcrições, por exemplo, são anotações na linha de glosa sintática XII (marcador de campo Toolbox, começando com a barra invertida aqui em fonte contornada: \SXII), e as traduções são exemplos diferentes de dados na linha XIII (\SXIIIe, para inglês, e \SXIIIn1, para o primeiro rascunho da tradução na linha nacional, o português).

Para algumas sessões adicionaremos outras anotações. Primeiro incluímos (semi-automaticamente) as informações extraídas das glosas morfológicas, como mostrado abaixo.

```
\ref      0002.690
\per      mawalaja
\SXII     jatätsu jatä ozoporywyt:
\SI       1       2       3           4
\SVI      jatä    tsu     jatä        ozoporywyt
\lx       jatä    tsu     jatä        porywyt
\SVII     dpron   pp      part        n
\SVIII    Unm_Nf  Unm_Pf  Unm_Pf      N_13 Unm_Ntense
\SIXn     este    como    é.que       costume
\SIXe     this    like    is.it.that  tradition
\CSIXn    este    como    é.que       nosso_costume
\CSIXe    this    like    is.it.that  our_tradition
\SXIIIn1  assim é que é nossa tradição:
\SXIIIe   it is like this that our tradition is:
\nts
\dt       02/May/2002
```

Palavras ortográficas na linha \SXII são divididas em unidades básicas sintáticas (por vezes chamadas palavras gramaticais, incluindo clíticos) na linha \SVI; aqui a posposição clítica *tsu* é separada do pronome governado *jatã*. Na versão anterior da ortografia do Aweti usada aqui na linha \SXII, posposições não constituíam palavras ortográficas. Os números na linha \SI foram adicionados à mão, mas o alinhamento com as colunas é sempre automático.

Para cada palavra, a forma de citação é dada na linha \lx (esse é um novo traço para ser incluído no AG) – compare a forma de citação ***porywyt*** 'tradição / costume / cultura' para a forma flexionada ***ozoporywyt***. Esta última pertence às categorias de forma 'Primeira Pessoa Plural Exclusiva Nominal' e 'Não Específica para Tempo Nominal', abreviadas na linha \SVIII por N_13 e Unm_Ntense.

Nas linhas seguintes temos a glosa para o significado lexical das palavras individuais (palavras funcionais teriam entradas de um tipo diferente aqui), e a tradução em português (\SIXn) e em inglês (\SIXe). As linhas \CSIXn e \CSIXe são adições ao AG, elas contêm a glosa completa, isto é, uma glosa que inclui efeitos de flexão (aqui somente relevante no caso de *ozoporywyt*).

A maior parte dessa informação foi adicionada semi-automaticamente, mas no caso de *jatã*, que funciona algumas vezes como um pronome demonstrativo e outras vezes como uma partícula de topicalização, nós tivemos que dissolver essa ambigüidade. A informação é obtida do banco de dados de glosagens morfológicas, como demonstrado abaixo na entrada para *ozoporywyt*, a única palavra morfologicamente complexa na sentença acima:

```
\MXII     ozoporywyt
\lx       porywyt
\MXIIIn   costume
\MXIIIe   tradition
\MI       1       2
\MVI      ozo-    porywyt
```

```
\MVII      f:poss-  n
\MVIII     Aff-     Unm
\MIXn      13-      costume
\MIXn      13-      tradition
\nts
\SVII      n
\SVIII     n_13 Unm_Ntense
\CMXIIIn   nosso_costume
\CMXIIIe   our_tradition
\dt        02/May/2002
```

Em princípio, este banco de dados é para glosagens morfológicas no formato AG. No entanto, nas entradas, nós encontramos não somente campos com dados morfológicos (os nomes de campo começam com \M), mas também campos que servem para o processo de interlinearização como um recurso para as tabelas de glosa sintática (nomes desses campos começam com \S, além da linha \lx); além disso, encontramos os campos usuais tais como o campo da data \dt e um campo para comentários \nts. As glosas de palavras (não-AG), incluindo efeitos de flexões, são aqui armazenadas nos campos com os rótulos \CMXIIIm e \CMXIIIe.

No começo a inserção automatizada de glosas é lenta, já que cada entrada (glosagem) morfológica para qualquer palavra em qualquer sentença tem que ser criada; mais tarde, quando as palavras mais freqüentes já têm suas glosagens morfológicas inseridas, o processo de anotação é mais rápido, mas mesmo assim tem que se reconhecer que ele nunca será tão rápido como a interlinearização morfêmica pura como a feita pela configuração padrão do **Toolbox**, que resulta em anotações no formato IMT tradicional. Na nossa configuração mais informação deveria ser entrada. Atualmente, para um minuto de texto gravado nós calculamos 40 minutos de anotação / inserção de glosagens (isto é, trabalho de interlinearização).

Uma parte da informação necessária nas tabelas de glosagens morfológicas por sua vez é obtida semi-automaticamente através de uma busca na informação das entradas relevantes nos bancos de dados lexicais. De novo, se uma entrada correspondente ainda não existe, ela poderia ser criada; desse modo o léxico aumenta com a anotação de textos (as glosas morfológicas funcionando como um nível intermediário entre o texto e o léxico).

Por razões técnicas, montamos três bancos de dados lexicais; um é para afixos e outro para (raízes de) palavras simples (incluindo partículas, clíticos e os chamados morfemas livres) – estes dois são usados no processo de interlinearização. O terceiro é para palavras complexas (derivadas ou compostas). Usando o método do "*jumping*" do **Toolbox**, se pode 'pular' diretamente de qualquer palavra num campo \lx (na glosagem morfológica ou sintática) para a entrada lexical correspondente, seja esta entrada no banco de dados para a palavra simples ou para palavras complexas.

Mostro aqui uma porção da entrada lexical para *porywyt*:

```
\lx   porywyt
\lc   [kaj]porywyt
\ps   n
\gn   costume
\ge   tradition
\dn   (um elemento da) cultura tradicional, os costumes antigos
\de   (an element of) traditional culture, the old costumes
\nq   obligatorily possessed?
\st   check
\simp S?
```

```
\crt   12/Set/2001
\dt    02/May/2002
```

A estrutura de entradas lexicais é um problema relativamente complexo que não pode ser discutido aqui; veja, por exemplo, Wittenburg, Peters e Drude (2002) e Wittenburg (2001) para detalhes. Usamos uma estrutura **MDF** (algo modificada ou estendida) para ser convertida em dicionários impressos através do **Multiple Dictionary Formatter** fornecido com o **Toolbox**.

Como demonstrado por Peter Austin (2002), o **Toolbox** pode ser usado para manter e ligar vários outros bancos de dados relevantes para qualquer projeto de documentação lingüística. Como um exemplo, para ser consistente, todos os campos que contêm abreviações usam um 'vocabulário controlado' (*range sets* na terminologia do **Toolbox**), e para cada abreviação criamos uma entrada num outro banco de dados. Em essas entradas providenciamos não somente a forma longa, mas uma descrição explícita sobre o tipo de entidade que é designada pelo termo. Eventualmente este banco de dados pode ser relacionado a outros sistemas terminológicos tais como a ontologia GOLD mencionada acima (Farrar e Langendon 2003).

Confira a entrada para Unm_Ntense, uma abreviação ocorrendo na linha \SVIII nas glosagens morfológica e sintática acima. A entrada inclui campos para a abreviação [\abrv] e o domínio ontológico [\dom] (isto é, se é uma categoria sintática ou morfológica ou uma outra entidade, no caso de uma categoria, se é uma categoria lexical – por exemplo, uma parte do discurso – ou uma categoria de formas – como um tempo, caso etc.); o nome extenso do termo (da categoria) [\long]; o status teórico [\type] (se é parte do sistema da língua ao nível morfológico / sintático, ou p.ex. uma classe fonológica ou semântica etc); uma explicação / descrição [\expl]; uma amostra [\spl] (aqui vazio); e campos organizacionais.

Obviamente os detalhes dos valores destes campos dependem da teoria lingüística subjacente a qual deveria ser explicitada.

```
\abrv   Unm_NTense
\dom    Syntax:Form
\long   Unmarked for nominal tense
\type   SUO (Syntactic Unit Ordering)
\expl   It is still unclear if nominal forms with the suffixes
        -(p)ut and -(z)an should be treated as derived or as
        inflected.  If inflected, the relevant categories are
        called Nominal Past and Nominal Future.  Most other
        noun forms are, then, unmarked for nominal tense.
\spl
\crt    11/Jun/2001
\dt     02/May/2002
```

As tabelas de glosagens sintáticas e as entradas morfológicas correspondentes podem ser completadas de acordo com o formato AG. Como dito acima, no projeto Aweti isso será feito somente para uma parte menor do corpus. Já mostrei uma tabela de glosagem sintática completa no **Toolbox** em Drude (2002) e não preciso repetir isso aqui. Acrescenta-se especialmente informação fonética e fonológica e informação sobre a estrutura gramatical e as relações entre os constituintes da sentença.

Poderíamos indicar, por exemplo, que as palavras 1 e 2 na tabela de glosa sintática acima constituem um grupo posposicional ('*postpositional frase*'). Formas analíticas de palavras, e em particular, constituintes descontínuos, deveriam de fato ser indicados imediatamente que forem identificados. Muitos outros detalhes teriam que ser

inseridos com pontos de interrogação indicando dúvida. Isto é previsto pelo formato AG.

Ainda não iniciamos essa fase em uma parte substancial dos dados, além para fins de testar e demonstrar os objetivos. Quanto à informação sobre a estrutura sintática e as relações, o preenchimento correto das linhas correspondentes pressupõe uma análise lingüística que em muitos casos ainda não foi concluída; por essa e outras razões, espera-se que essa fase seja lenta e demorada e isso é a razão pela qual somente uma parte muito pequena do corpus será completada desse modo. Mas isso não atinge a validade e utilidade do formato AG como um todo – ao contrário, o formato prevê explicitamente que a informação será preenchida em etapas diferentes.

Com estes comentários fechamos nossa descrição da metodologia adotada no Projeto de Documentação da Língua Aweti. Esperamos que alguns métodos sejam úteis para outros projetos semelhantes. A riqueza lingüística brasileira merece ser documentada, pois ela corre o risco de diminuir drasticamente nas próximas gerações.

Referências

Audacity (*computer program*) http://audacity.sourceforge.net/

AUSTIN, Peter K. "Developing Interactive Knowledgebases for Australian Aboriginal Languages – Malyangapa". Palestra no *Workshop on Australian Aboriginal Languages*, University of Melbourne, March 2002.
http://www.linguistics.unimelb.edu.au/contact/staff/peter/Malyangapa.pdf

BAKKER Dik, DAHL, Oesten, LEHMANN, Christian, e Siewierska, Anna. "Eurotyp guidelines. Technical report". Fondation Europeenne de la Science, Strassbourg. (EUROTYP Working Papers). 1994.

BOW, Cathy, HUGHES, Baden e BIRD, Steven. "Towards a General Model of Interlinear Text". Palestra na *third EMELD Conference on Digitizing & Annotating Texts & Field Recordings*. LSA Institute, Michigan State University, July 11th -13th. 2003.
http://saussure.linguistlist.org/cfdocs/emeld/workshop/2003/bowbadenbird-paper.html

CROFT, William. "Abbreviations and symbols for interlinear morpheme translation". Em: *Typology and universals*, 2ª edição. 2003.
http://lings.ln.man.ac.uk/Info/staff/WAC/Papers/TypAbbrev.pdf

DALBY, Andrew. *Language in Danger: The Loss of Linguistic Diversity and the Threat to our Future*. New York: Columbia University Press. 2003.

DOBES: http://www.mpi.nl/DOBES
http://www.mpi.nl/DOBES/teams/Aweti/Aweti.html
http://www.mpi.nl/DOBES/teams/Kuikuro/Kuikuro.html
http://www.mpi.nl/DOBES/teams/Trumai/Trumai.html
http://www.mpi.nl/DOBES/teams/TIDEL/TIDEL.html

DRUDE, Sebastian. "Advanced Glossing – a language documentation format and its implementation with Shoebox". Trabalho apresentado no *LREC-Workshop* em Las Palmas, May 2002.
http://www.mpi.nl/DOBES/meetings/lrec2002/lrecWorkshop.pdf

DRUDE, Sebastian. “Digitizing and Annotating Texts and Field Recordings in the Awetí Project”. Palestra na *third EMELD Conference on Digitizing & Annotating Texts & Field Recordings*. LSA Institute, Michigan State University, July 11th -13th. 2003.
http://saussure.linguistlist.org/cfdocs/emeld/workshop/2003/drude-paper.html

FARRAR, Scott e LANGENDOEN, D. Terence. “Markup and the GOLD Ontology”. Palestra na *third EMELD Conference on Digitizing & Annotating Texts & Field Recordings*. LSA Institute, Michigan State University, July 11th -13th. 2003.
http://saussure.linguistlist.org/cfdocs/emeld/workshop/2003/paper-terry.html ver também
http://www.linguistics-ontology.org

FRANCHETTO, Bruna. “O conhecimento científico das línguas indígenas da Amazônia no Brasil”. In: *As línguas amazônicas hoje*, ed. por QUEIXALÓS, Francisco e RENAULT-LESCURE, O., p. 165-82. São Paolo. 2000.

GORDON, JR., Raymond G. (ed.). *Ethnologue: Languages of the World, Fifteenth edition*. Dallas, Texas: SIL International. 2005. http://www.ethnologue.com/

HIMMELMANN, Nikolaus P. “Documentary and descriptive linguistics”. *Linguistics*, 36.1, p.161-195. 1998.

HOCKETT, Charles F. “Two models of grammatical description”. *Word*, 10. p. 210–234. 1958.

IMDI: ISLE Metadata Initiative. “Metadata Elements for Session Descriptions”. Draft Proposal Version 3.0.3, July 2003.
http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.3.pdf
Sobre ISLE ver também: http://www.ldc.upenn.edu/sb/isle.html

KRAUSS, Michael. “The Worlds Languages in Crisis”. In: *Language*, 68. p. 4-10. 1992.

LEHMANN, Christian. “Directions for interlinear morphemic translations”. *Folia Linguistica* (Acta Societatis Linguisticae Europaeae), XVI: 199-224. 1982.

LIEB, Hans-Heinrich, DWYER, Arienne M. e ANDERSON, Gregory D. “Approaches to Morphosyntactic Annotation”. DOBES internal Working Paper. 2001.
http://www.linguistlist.org/~workshop/markup/DOBES-markup.html

LIEB, Hans-Heinrich e DRUDE, Sebastian. “Advanced Glossing: A language documentation format. DOBES internal Working Paper. (1ª versão; uma segunda, mais explicativa está sendo preparada). 2000.
http://www.mpi.nl/DOBES/applicants/Advanced-Glossing1.pdf

MOORE, Denny. “Endangered languages of Lowland tropical South América”. In: *Endangered Languages and Language Documentation*, BRENZINGER, Mathias (ed.). Em preparação.

MOSELEY, Christopher, Ed. *Encyclopedia of the World's Endangered Languages*. Richmond: Curzon, Curzon Language Family Series. 2001.

MPI Tools: **ELAN**, **Econv**, o **IMDI editor** e **IMDI browser**, entre outros, estão sendo desenvolvido pela equipe “TIDEL” no Max Planck Institute in Nijmegen.
http://www.mpi.nl/tools

NETTLE, Daniel e ROMAINE, Suzanne. *Vanishing voices: The Extinction of the World's Languages*. Oxford: Oxford University Press. 2000.

RODRIGUES, Aryon Dall'igna. "Línguas Indígenas: 500 anos de descobertas e perdas". Em: *Ciência Hoje*, 16/No 95. p. 20-26. 1993.

Sound Forge XP (*computer program*): http://www.sonicfoundry.com/Products/showproduct.asp?PID=668

Toolbox/Shoebox (*computer programs*): www.sil.org/computing/toolbox/

Transcriber (*computer program*): http://www.etca.fr/CTA/gip/Projets/Transcriber/

WITTENBURG, Peter. "Survey of lexical structures". DOBES internal working paper. Cf. Wittenburg et.al. 2002) 2001.

WITTENBURG, Peter. "Effects of Compression on Linguistically Relevant Speech Analysis Parameters". (Internal document presented to the DOBES teams, among others.) 2002.

WITTENBURG, Peter, PETERS, Wim, e DRUDE, Sebastian. "On lexical structures in language engineering and field linguistics". Palestra na *LREC-Conference* em Las Palmas, May 2002. http://www.mpi.nl/DOBES/meetings/lrec2002/lrecWorkshop.pdf