PUL∅ - Para um sistema de tradução semi-automática português-libras

Ronaldo Martins[1,2], Jorge Pelizzoni[2], Ricardo Hasegawa[2]

[1]Faculdade de Filosofia, Letras e Educação – Universidade Presbiteriana Mackenzie
Rua da Consolação, 930 – 01302-907 – São Paulo – SP – Brazil

[2]Núcleo Interinstitucional de Lingüística Computacional (NILC)
Av. Trabalhador São-Carlense, 400 – 13560-970 – São Carlos – SP – Brazil
ronaldomartins@mackenzie.com.br, {jorgemp, rh}@icmc.usp.br

Abstract. *This paper describes a prototype for an interlingua-based human-aided machine translation system from Brazilian Portuguese into a linearized script of Libras, the Brazilian Sign Language. It is the first module of an experimental system for cross-modal translation aiming at overcoming the digital divide that segregates the Brazilian deaf community. The system is also expected to be used as a Brazilian Portuguese learning tool for Libras native speakers.*

Resumo. *Este artigo tem por objetivo descrever o protótipo de um sistema de tradução automática auxiliada por humanos, baseado em interlíngua, entre o português brasileiro e uma representação linearizada de libras, a língua brasileira de sinais, utilizada pela comunidade dos surdos do Brasil. Trata-se do primeiro módulo de um sistema experimental de tradução intermodal entre uma língua oral-auditiva (o português brasileiro) e uma língua gestual-visual (libras), cujo objetivo último está relacionado à inclusão digital das comunidades de surdos brasileiros e ao desenvolvimento de ferramentas pedagógicas de ensino de língua portuguesa para usuários de línguas de sinais.*

1. Introdução

O presente artigo tem por objetivo reportar o desenvolvimento do PUL∅ – Portuguese-UNL-LIST deOralizer –, versão experimental de um sistema de tradução automática unidirecional de uma língua oral-auditiva, o português, para a representação linear (*LIbras Script for Translation – LIST*) de uma língua gestual-visual, a língua brasileira de sinais, ou libras. O principal objetivo do PUL∅ é converter uma sentença originalmente produzida em língua portuguesa para uma transcrição especializada de libras que seja um compromisso entre (1) simplificação do processo de tradução e (2) suficiência para a síntese de sinais em libras.

O PUL∅ acompanha a tecnologia de sistemas de tradução automática auxiliada por humanos [Hutchings and Sommers 1992], na medida em que exige, no processo de tradução, alguma interação com o usuário humano proficiente em língua portuguesa. Não é objetivo da ferramenta realizar toda a conversão português-LIST de forma completamente automática, mas contar com o conhecimento humano – particularmente com o conhecimento de difícil ou ainda impossível formalização computacional – para resolver as ambigüidades e os desvios lingüísticos que possam ser observados nas sentenças de entrada

do sistema. No entanto, a expectativa da ferramenta é a de que seu usuário final possa ser monoglota e não-especialista, ou seja, de que possa constituir mão-de-obra mais barata e menos qualificada do que a requerida em uma atividade de tradução comum. Principalmente, a ferramenta parte do compromisso de não exigir desse usuário final nenhum conhecimento de libras ou de qualquer outra língua gestual-visual.

O PUL∅ toma, como entrada, uma subvariedade simplificada da língua portuguesa (aqui chamada "português normalizado"), desprovida de elipses, topicalizações, anacolutos, anáforas, ambigüidades léxicas e sintáticas e outros acidentes lógico-gramaticais que pudessem vir a afetar o desempenho da ferramenta. Este processo de normalização corresponde exatamente à parte interativa da ferramenta e segue, em linhas gerais, as estratégias disponíveis para a produção de "línguas controladas" [Wojcik and Hoard 1995]. Para a análise semântica desse subconjunto do português, utilizou-se a UNL, ou *Universal Networking Language* [Uchida et al. 1999], uma linguagem de representação do conhecimento, que operou, no protótipo, como interlíngua, para a qual era convertida a sentença em língua portuguesa, e da qual era gerada a representação linearizada da língua brasileira de sinais (LIST). A representação UNL ofereceu a perspectiva de desenvolvimento de um sistema de tradução baseado, principalmente, em informação de natureza semântica, em detrimento das estruturas sintáticas, que são consideravelmente diferentes entre português e libras. Em última instância, esta escolha representou a adoção de técnicas de transdução de informação originalmente empacotada em estruturas lineares (a sentença do português), que foi inicialmente convertida para estruturas reticuladas (os grafos UNL) e, em seguida, reapresentada como nova estrutura linear (as listas de LIST).

A ausência de recursos e de tecnologia previamente disponíveis e o curto prazo de tempo para a implementação do protótipo fizeram que a prototipagem do PUL∅ se visse limitada a uma única história em quadrinhos da Turma da Mônica, selecionada pela coordenação geral do projeto, e composta de 12 frases, de complexidade média. Embora este corpus possa parecer excessivamente limitado, deve-se salientar que nosso objetivo primeiro envolvia apenas a verificação da viabilidade da abordagem escolhida, e não o desenvolvimento de uma ferramenta genérica, robusta, de aplicabilidade ilimitada. As conclusões parciais extraídas desta experiência, embora remetam diretamente aos fatos lingüísticos e problemas computacionais associados a esse pequeno conjunto de 12 frases, permitiram a antecipação de muitos outros problemas e a revisão de algumas estratégias adotadas, que se espera possam ser testadas e eventualmente referendadas para um conjunto maior de dados, em uma desejável segunda etapa do projeto.

Durante o desenvolvimento do protótipo, inúmeros recursos lingüísticos foram desenvolvidos. São eles: a notação LIST, o formalismo gramatical NL-UNL, a gramática português-UNL, o dicionário português-UNL, a gramática UNL-LIST e o dicionário UNL-LIST. Paralelamente foram também produzidos subsistemas computacionais que atuam de forma modular na plataforma PUL∅: o normalizador do português (Kaapor), a ferramenta de análise genérica NL-UNL (HERMETO) e o conversor UNL-LIST (ManateCo), construído a partir dos aplicativos de geração (DeConverter, versão 2.5) e de construção de dicionários (DicBld, versão 2.1) fornecidos pela *Universal Networking Digital Language Foundation*, de Genebra. O principal objetivo deste artigo, a par da apresentação geral da abordagem adotada, consiste em uma rápida exposição de cada um desses módulos e recursos, bem como das escolhas e decisões técnicas a eles associadas.

2. Tradução Intersemiótica Automática

A atividade prevista no projeto global – a recodificação de uma mensagem originalmente produzida no registro da escrita de uma linguagem verbal natural (o português brasileiro) para uma linguagem gestual-visual (libras) - enquadra-se no que vem sendo chamado, nos estudos da linguagem, de tradução intersemiótica ou, mais especificamente, *modality translation* [Jakobson 1949, Santaella 2001, Zimmermann and Vanderheiden 2001]. Trata-se, até onde pudemos observar, de um domínio já razoavelmente explorado dentro dos estudos da tradução automática, para o qual não foram encontradas, no entanto, outras iniciativas para a língua portuguesa ou para a língua brasileira de sinais. A bibliografia recenseada, principalmente relativa a sistemas de tradução do inglês para a *American Sign Language*, parece convergir para três pontos principais: (1) os sistemas de tradução automática intermodal acompanham, em linhas gerais, os princípios, abordagens e técnicas já desenvolvidos para os sistemas intramodais (de uma língua oral-auditiva para outra língua oral-auditiva), a despeito das diferenças de suporte; (2) os sistemas de tradução intermodal se subdividem, na verdade, em dois subsistemas: o de tradução de uma língua oral-auditiva para um sistema de escrita da língua gestual-visual; e o de síntese de sinais (gestual-visuais) a partir desse sistema de escrita; e (3) a complexidade da tarefa está evidentemente relacionada ao sistema de escrita da língua gestual-visual adotado. Nesta seção, cada um desses tópicos será referido separadamente, para que se possam estabelecer os fundamentos teóricos para as escolhas que foram feitas.

Em primeiro lugar, cabe referir a idéia de que não há diferenças expressivas entre as tecnologias de tradução intermodal e as tecnologias de tradução intramodal. Verifica-se, lá como cá, a existência de pelo menos três abordagens predominantes [Dorr et al. 1999]: a tradução baseada exclusivamente em conhecimento lingüístico, ou seja, em dicionários e gramáticas (*Language-Based Machine Translation* – LBMT); a tradução baseada em conhecimento, ou seja, em dicionários, gramáticas e, adicionalmente, enciclopédias e bases de conhecimento (*Knowledge-Based Machine Translation - KBMT*); e a tradução baseada em exemplos, ou seja, em dicionários, gramáticas e *corpora (Example-Based Machine Translation – EBMT)*. Os dois primeiros casos constituiriam, principalmente, modelos de tradução baseada em regras, ou na explicitação do conhecimento lingüístico inato do falante; o último seria particularmente amparado em análises e dados estatísticos. O primeiro modelo, em função do custo relativamente mais baixo se comparado aos demais, seria mais adequado para sistemas mais genéricos e mais robustos, mas produziria resultados menos satisfatórios e mais sujeitos a erro. Os dois últimos, por envolverem o desenvolvimento de recursos mais dispendiosos (enciclopédias e *corpora* convenientemente anotados, separados por domínio do conhecimento), produziriam resultados mais exatos, mas seriam indicados apenas para sistemas mais especializados, de domínio restrito.

Do ponto de vista da técnica, são citadas duas linhas principais: a tradução direta e a tradução indireta. A tradução direta prevê, em linhas gerais, que a língua-alvo seja considerada o próprio instrumento de análise da língua-fonte. A tradução indireta prevê o desenvolvimento de uma forma de representação intermediária entre a língua-fonte e a língua-alvo. Esta forma de representação pode ser dependente das línguas envolvidas, no sentido de constituir uma interface específica (unidirecional ou bidirecional), ou pode ser independente tanto da língua-fonte quanto da língua-alvo, procurando organizar-se como uma outra língua, artificial, autônoma, neutra, porém mais adequada ao processamento

automático (porque livre de ambigüidade, por exemplo). No primeiro caso, fala-se em tradução indireta baseada em transferência; no segundo, em tradução indireta baseada em interlíngua.

O PUL∅ foi deliberadamente proposto como um sistema de tradução automática a) baseado exclusivamente em conhecimento lingüístico e b) que utiliza a estratégia de tradução indireta por interlíngua. A primeira opção, caracterizada pelo fato de o sistema prever apenas o desenvolvimento de dicionários e gramáticas e dispensar a construção de outros repositórios de informação, se explica, em larga medida, por uma série de restrições operacionais que diziam respeito, particularmente, à definição do *corpus* a ser trabalhado. Na medida em que se definiu que se trabalharia sobre histórias em quadrinhos, de temática bastante variada, com a predominância de gêneros primários do discurso [Bakthin 1953], não se revelou viável a construção de uma base de conhecimento que permitisse o equacionamento de todas as ambigüidades relativas aos textos de entrada. Por este motivo, decidiu-se, desde o início, que o sistema proveria essas informações por meio da interação com o usuário humano, que faria as vezes, portanto, de repositório adicional de informações, substituindo enciclopédias e outras bases de conhecimento normalmente adotadas em outros sistemas de tradução. A adoção de uma estratégia de tradução indireta baseada em interlíngua foi principalmente derivada da experiência prévia do grupo com um modelo desta natureza (o sistema UNL); da aposta nas virtudes desta proposta, particularmente no que tange à facilidade de expansão do modelo, com a incorporação de outras línguas, o que permitiria que pudéssemos desenvolver, no futuro próximo, uma plataforma multilíngüe de tradução para libras; e da verificação das diferenças – semânticas, principalmente – entre o português e libras, particularmente quando se percebeu que a organização fonológica e morfossintática da língua de sinais seria afetada por um conjunto de marcadores semânticos (de natureza visual) que normalmente não está disponível nas descrições gramaticais das línguas naturais. Desta forma, previu-se que a tradução português-libras mereceria contar com um nível de representação intermediário, que fosse suficientemente autônomo, seja em relação à língua-fonte, seja em relação à língua-destino.

A modularização dos sistemas de tradução intermodal parece ser uma opção mais pacífica do que a determinação de suas abordagens e estratégias. Quando se confrontam os dois diferentes suportes – bidimensional, no caso da língua oral-auditiva, e tridimensional, no caso da língua gestual-visual – parece ser uma escolha razoavelmente óbvia a subdivisão do processo em duas etapas muito distintas: (1) tradução da língua oral-auditiva para algum tipo de transcrição simplificada da língua gestual-visual e (2) síntese de fala (no sentido lato), em que a transcrição é "executada", ou seja, convertida num enunciado gestual-visual sintetizado. No caso do projeto em tela, estas duas etapas foram, desde o início, distribuídas entre atores diferentes. Entre os dois módulos, no entanto, observou-se, desde sempre, a necessidade da compatibilidade das informações: a saída provida pelo tradutor deveria obrigatoriamente respeitar as necessidades estabelecidas pelo sintetizador. No entanto, dada a falta de sincronia entre os desenvolvedores das duas partes do sistema, particularmente porque o trabalho de análise teria precedido bastante a contratação de uma equipe de síntese, acabou-se por definir unilateralmente um formato de saída, inspirado na notação linear proposta por [Felipe 1998], modificada porém para eliminar as ambigüidades estruturais e os problemas de digitação derivados desta proposta. Esta nova notação – batizada de LIST – será apresentada na Seção 3.

A arquitetura geral do PUL∅ é apresentada na Figura 1.

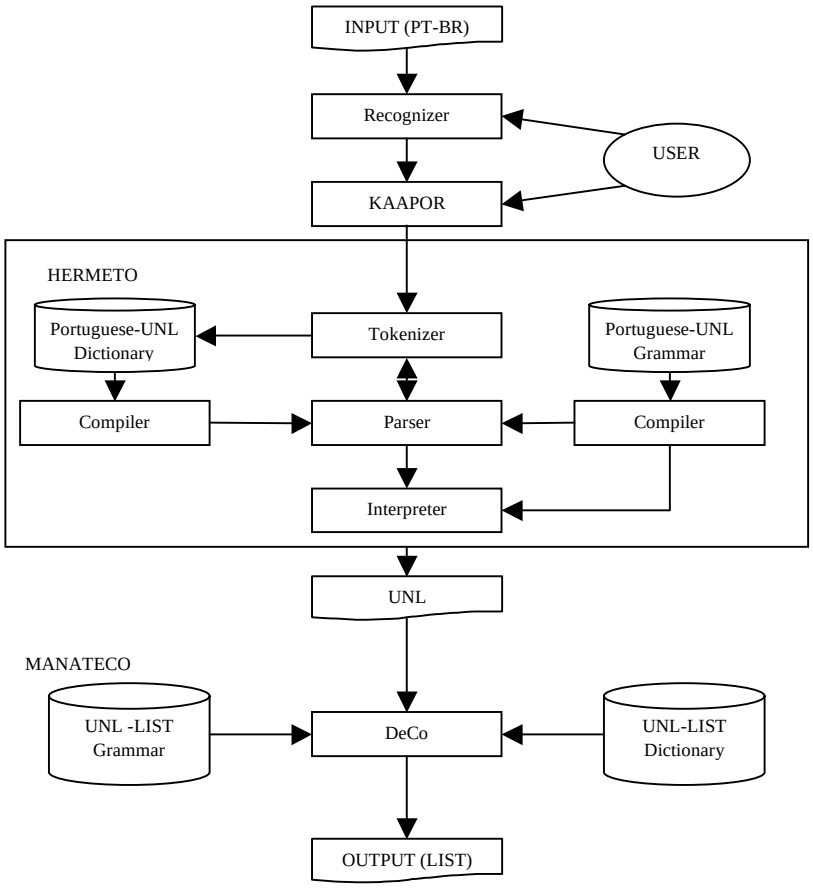

Figura 1. Arquitetura Geral do PUL∅

O dado de entrada – a fala de uma personagem em cada um dos quadrinhos de uma história da Turma da Mônica – é digitado pelo usuário e é desmembrado automaticamente, pelo *recognizer*, em unidades de processamento (sentenças), a partir de algumas fronteiras sentenciais pré-estabelecidas (ponto, ponto de interrogação, ponto de exclamação, etc). Em seguida, a sentença isolada é normalizada, pelo Kaapor, semi-automaticamente, com a interação do usuário, para que se possam resgatar as informações não disponíveis em sua estrutura superficial (elipses, referências anafóricas, etc.) e para que se possam resolver todas as ambigüidades observadas. O *normalizer* também verifica a consistência ortográfica do vocabulário utilizado, reconhecendo interjeições e grafias desviantes do padrão da língua ("*obrigadooooo...*", por exemplo), muito comuns na reprodução da fala das personagens, principalmente do Cebolinha, que troca sistematicamente os "r" pelos "l". A sentença normalizada é subdividida em itens lexicais, pelo *tokenizer*, que recolhe as informações relativas a cada entrada no dicionário e as repassa para o *parser*, que opera a análise sintática automática a partir das possibilidades combinatórias contidas na gramática português-UNL. As regras sintáticas estão associadas a regras de projeção semântica, ativadas pelo *interpreter*, que gera finalmente o grafo UNL para a sentença analisada. A representação UNL, um conjunto de relações binárias unidirecionais entre itens do vocabulário da linguagem UNL, serve de entrada para o DeCo, ferramenta genérica de decodificação de UNL, que converte a representação reticulada em uma nova lista, agora em LIST, com o auxílio da gramática e do dicionário UNL-LIST. Este resultado final – a sentença em LIST

– serviria de entrada ao segundo módulo do tradutor geral português-libras, responsável pela síntese de fala.

3. LIST - LIbras Script for Translation

Para a implementação do sistema, fez-se necessária uma transcrição (linear) de libras (1) afeita ao tratamento computacional, (2) concebida especialmente para fins de tradução (semi)automática intermodal e (3) que fosse um compromisso entre simplificação do processo de tradução e do desenvolvimento do tradutor e suficiência para a *síntese de sinais* na língua brasileira de sinais. Mais especificamente, a formalização proposta constituiu o formato de saída de PUL∅ e o formato de entrada do sintetizador de sinais. Seu objetivo seria representar, de forma não ambígua, todas e apenas as informações necessárias para a geração de sinais de libras a partir de uma sentença em língua portuguesa, considerando, particularmente: a) que português e libras são línguas completamente distintas, não apenas em relação à estrutura (gramática e léxico), mas ao próprio suporte; b) que a sentença em língua portuguesa possui redundâncias (como a concordância de número, de gênero e de pessoa gramatical) que seriam suprimidas no processo de representação em libras; c) que a sentença em língua portuguesa possui informações (como a ordem dos constituintes) que nem sempre são relevantes para a representação em libras; d) que a sentença em língua portuguesa possui lacunas (informação sobre classificadores, elipses, anáforas e outras pró-formas, nominais e verbais) que deveriam ser preenchidas, ainda que de forma arbitrária, para que se pudesse assegurar a gramaticalidade e a semanticidade dos enunciados de libras.

O vocabulário LIST é constituído de *LIST Words* (*LWs*), que corresponderiam à representação linear (bidimensional) dos sinais de libras. Por razões mnemônicas, este vocabulário foi definido a partir do conjunto de itens lexicais de língua portuguesa, ao invés de procurar representar índices para a síntese de fala, como propõem muitos outros sistemas de escrita para línguas de sinais (HamNoSys, por exemplo), de aplicação diversa à pretendida para LIST, no entanto. Cada LW é composta de uma *headword*, necessariamente, seguida de uma matriz atributo-valor, opcionalmente.

Como o próprio nome sugere, LIST tenta, na medida do possível, aproximar-se de uma estrutura linear, uma simples lista de sinais a serem executados em seqüência, separados por espaços em branco. Entretanto, introduzimos alguns “operadores/estruturadores frasais” para melhor dar conta de alguns fenômenos de libras, como o paralelismo no processo de geração de sinais (que pode mobilizar, simultaneamente, a ação das mãos e a alteração dos sinais da face). e os grupos rítmicos (dada a concentração do processo de geração de sinais em estruturas "prosódicas" com intervalos de "silêncio" de duração variável). Por fim, cabe referir que a estrutura do documento LIST acompanha a estrutura de um documento XML, em que seriam utilizadas algumas etiquetas (*tags*) específicas.

4. Universal Networking Language (UNL)

Entre os objetivos assinalados para o desenvolvimento do protótipo, foram duas principalmente as questões que nos conduziram à adoção do modelo de tradução baseada em interlíngua: a) a idéia de que a representação LIST, proposta como saída do sistema, deveria ser autoconsistente, autônoma e independente em relação ao processo de tradução e síntese de sinais, de forma a permitir que as estratégias empregadas em ambos os processos pudessem ser substituídas sem que haja necessidade de alteração da representação e sem que

fossem inutilizados os documentos já codificados; e b) que seria desejável que a representação LIST fosse semelhante a outros formalismos computacionais de representação do conhecimento, para que se pudesse assegurar a intercambialidade entre os códigos e o desenvolvimento de filtros que permitissem, à LIST, servir de representação-alvo de outros sistemas de análise sintático-semântica (do português e também de outras línguas).

Estes dois compromissos, aliados à perspectiva de ampliar o sistema, transformando-o em uma plataforma efetivamente multilíngüe, permitiram perceber que seria interessante incorporar, como estratégia de tradução, uma representação intermediária, supostamente não-marcada e independente em relação à língua-fonte e à língua-alvo, para que se pudesse permitir uma maior intercambialidade de informações, e para assegurar portabilidade e robustez a todo o sistema. Entre os vários formalismos disponíveis, optou-se pela *Universal Networking Language*, tendo em vista: a) a experiência anterior do grupo com este modelo de formalização; b) o seu caráter efetivamente plurilingüístico, dado que a iniciativa envolve grupos lingüísticos muito mais variados do que as outras abordagens disponíveis; c) o seu caráter público, na medida em que as patentes associadas pertenceriam à ONU; e d) sua abrangência, já que a flexibilidade da representação permitiria à UNL contemplar as informações de natureza visual freqüentemente expurgadas dos outros modelos, de base estritamente oral-auditiva.

Infelizmente, não cabe neste texto uma descrição mais pormenorizada da UNL, e de sua estratégia de representação do conhecimento semântico veiculado pela sentença por meio de grafos cujos nós representariam conceitos universais e cujos arcos [entre nós] constituiriam relações semânticas binárias unidirecionais entre conceitos. Para maiores informações, consulte-se [Uchida et al. 1999] ou o sítio da UNDL Foundation, em Genebra (http://www.undl.org).

5. Corpus

Para o desenvolvimento do protótipo, decidiu-se pelo trabalho em um *corpus* constituído, inicialmente, por histórias infantis. Duas histórias chegaram a ser codificadas, mas se percebeu, muito cedo, que as histórias infantis escritas em língua portuguesa, quando recontadas em libras, envolveriam adaptações e alterações que muito dificilmente poderiam ser replicadas em um ambiente computacional com a tecnologia disponível. A presença freqüente de um vocabulário diferenciado (e infantilizado), o revezamento entre discurso direto e discurso indireto, e o apoio indispensável nas ilustrações, com as quais a representação em sinais deveria competir, indicaram que, nessa primeira etapa, seria mais prudente trabalharmos com um corpus de outra natureza, em que algumas dessas variáveis pudessem ser mais bem controladas. Por este motivo, preferiu-se trabalhar com histórias em quadrinhos, porque se atingiria o mesmo público-alvo (as crianças surdas); porque se restringiria o gênero de trabalho apenas ao diálogo, ou discurso direto; porque se poderia criar uma interface mais atraente, em que os sinais correspondentes a cada balão seriam gerados no momento em que este fosse clicado pelo leitor, garantindo um procedimento de leitura simples, intuitivo e que respeitasse o ritmo de cada usuário. Entre as histórias em quadrinhos, optou-se pelas tirinhas semanais da Turma da Mônica, editadas pela Editora Globo, e disponíveis no Portal da Mônica, em http://www.portaldamonica.com.br.

Selecionou-se, inicialmente, uma só história – a de número 74 – que consistia em um panfleto instrutivo a respeito dos cuidados para se evitar a disseminação da dengue. As 12 sentenças da história foram codificadas para as representações LIST e UNL desejáveis (ou possíveis) para cada um dos enunciados apresentados. A partir dessas representações, foram produzidos os recursos lingüísticos (dicionários e gramáticas) necessários para que pudesse ser emulado computacionalmente o comportamento humano.

6. O pulo português-UNL

No processo de análise do português, para sua tradução em UNL, foram desenvolvidos vários recursos, lingüísticos e computacionais, descritos resumidamente nesta seção. Muitos desses recursos, embora tenham sido elaborados especificamente para o projeto em tela e para o *corpus* especificado, são customizáveis para outros projetos e para outros *corpora*.

O módulo Kaapor é o responsável pela normalização, ou seja, por realizar as correções necessárias no texto para que ele possa ser traduzido. Tal módulo tenta, na medida do possível, realizar as correções automaticamente, sem a intervenção do usuário; mas existem alguns casos em que o auxílio de um usuário falante da língua portuguesa é indispensável. Isso acontece, por exemplo, no tratamento dos desvios ortográficos, de algumas interjeições (de sentido ambíguo), no preenchimento das elipses e na recuperação das relações anafóricas. As histórias são originalmente apresentadas como uma figura (extensão *.gif*) e devem ter suas falas transpostas para um arquivo-texto (extensão *.txt*), na medida de um balão por linha. O módulo oferece ao usuário a opção de visualização da história em quadrinhos original. O processamento do texto se inicia a partir do arquivo-texto: a ferramenta de normalização começa pela análise/conversão das sentenças.

A gramática português-UNL, construída no âmbito deste projeto, pode ser definida pela sêxtupla <N, T, P, I, W, S> , em que: N = conjunto de símbolos não-terminais; T = conjunto de símbolos terminais; P = conjunto de regras de produção sintática; I = conjunto de regras de interpretação semântica; W = prioridade de aplicação das regras de produção sintática (inteiro de 0 a 255); S = símbolo inicial. Para efeito de simplificação, o conjunto de regras de produção sintática (P) e o de interpretação semântica (I) foram representados conjuntamente, como campos distintos de uma mesma regra. Assim, as regras da gramática possuem o formato: **p -> i,** em que $p \in P$, e $i \in I$. As regras (p) de produção sintática acompanham uma gramática livre-de-contexto, com indicação explícita de prioridade de aplicação de cada regra, definida pela fórmula: **a[w] := b,** em que $a \in N$, $b \in N \cup T$, e $w \in W$. Nessa fórmula, quanto maior o valor de 'w', tanto menor a prioridade de aplicação da regra (0 = prioridade máxima de aplicação). Se duas ou mais regras compartilharem a mesma prioridade, sua ordem de aplicação corresponderá à ordem das regras no arquivo da gramática. Para simplificação do número de regras e para restrições, foram utilizados vários operadores, usualmente empregados em outros formalismos gramaticais, como os de opcionalidade, de mútua exclusividade, etc.

O dicionário português-UNL utilizado pelo projeto correspondeu a um arquivo texto em que as entradas foram representadas, em cada linha, pela sua associação entre a entrada em língua portuguesa, caracterizada por formas livres (lexias simples, lexias compostas, lexias complexas) ou por formas presas (como raízes, afixos e partes de *collocations*); a representação semântica da entrada na rede UNL de conceitos; e um conjunto (não-

ordenado) de pares de atributo-valor, de natureza fonético-fonológica, morfossintática, semântica, pragmático-discursiva e outras, na medida das necessidades da aplicação.

O ambiente Hermeto compreende cinco diferentes submódulos: um tokenizer, o compilador do dicionário português-UNL, o compilador da gramática português-UNL, um analisador sintático automático, e um interpretador semântico automático. Ele toma, como ponto de partida, a sentença normalizada do português, o dicionário e a gramática português-UNL (em sua versão texto) e produz, como saída, o grafo UNL correspondente. Esta ferramenta foi desenvolvida para facilitar o trabalho e a depuração no desenvolvimento de qualquer *parser* para qualquer língua de origem. Portanto, ela permite tanto a criação e edição de uma gramática como a manipulação de um *corpus* e o uso deste *corpus* na depuração da gramática e dicionário. O uso da DLL que incorpora o tokenizer, o parser, o interpretador e a gramática já compilados, mais o dicionário no formato descrito acima, resulta em uma árvore sintática e no código UNL da frase de entrada normalizada. Assim, a DLL admite, como dado de entrada, apenas sentenças isoladas. Conseqüentemente, o Hermeto, embora não faça qualquer restrição quanto à dimensão, quanto à forma, ou quanto ao domínio da sentença de entrada, não é capaz de resgatar relações de natureza extra-sentencial, seja por referência ao co-texto imediato, seja por referência ao contexto situacional, incluídas as substituições, as elipses, as pronominalizações e todas as relações anafóricas.

7. O pulo UNL-LIST

O módulo de tradução UNL-LIST, ManateCo, é um dos itens mais prototípicos de PUL∅ exatamente por se basear fortemente no DeConverter, ferramenta genérica de decodificação UNL-NL fornecida pela *UNDL Foundation*, que apresenta algumas características que limitam ou desaconselham sua aplicação ao problema em vista, pelo menos em sua forma mais geral. O ManateCo propriamente dito não passa de um script Perl que (1) isola expressões UNL de arquivos-texto de entrada (de extensão *.unl*, necessariamente), devidamente marcadas com as etiquetas *{unl}* e *{/unl}*; (2) submete-as ao DeCo, devidamente alimentado com regras de tradução UNL-LIST e um dicionário UNL-LIST; (3) procede a uma formatação muito simples da resposta do DeCo; e (4) gera arquivos-texto de saída (de mesmo *path* que os de entrada, mas com extensão *.list*) idênticos aos de entrada, exceto pela substituição (comportamento *default*) das expressões UNL pelas respectivas versões LIST, marcadas com *{list}* e *{/list}*, ou pela justaposição (mediante a opção de linha de comando --*unl*) destas em seguida àquelas.

Apesar de sabermos com antecedência das limitações do DeCo, resolvemos utilizá-lo exatamente para levantar requisitos para o projeto de uma ferramenta de decodificação também genérica mas mais compatível com a tradução intermodal especificamente e, em geral, mais amigável, apoiando-se em algum formalismo de mais alto nível, mais declarativo.

8. Integração do sistema

Atualmente os módulos do PUL∅ se articulam da seguinte forma: o ManateCo funciona como uma aplicação independente que, dados pré-documentos LIST (i.e., documentos LIST que prescindam exatamente das versões LIST de suas expressões UNL), gera documentos

LIST correspondentes completos; o Hermeto está disponível como uma DLL exportando uma função que, dada uma sentença em português normalizado, gera a expressão UNL correspondente; e o Kaapor se apresenta como uma aplicação que, por conveniência, executa todo o procedimento de tradução de forma transparente (1) gerando o arquivo normalizado (extensão *.kpr*), (2) submetendo cada uma de suas sentenças à DLL Hermeto, (3) usando o resultado para gerar um pré-documento LIST (*.unl*) e (4 invocando o ManateCo para o arquivo em questão, o que resulta num documento LIST (*.list*) completo correspondente à historinha editada.

9. Avaliação global e perspectivas de trabalhos futuros

As seções anteriores deixam evidente a complexidade da tarefa executada nessa fase do projeto – a tradução de português para uma transcrição de libras suficiente para alimentar um sintetizador de fala. A estratégia de prototipação, focando uma única história até o momento, mostrou-se apropriada, embora pareça ser bastante restritiva, uma vez que em um corpus maior de sentenças em português deverão surgir outros problemas a serem atacados. No entanto, tomou-se o devido cuidado de se tratar fenômenos bastante freqüentes, de modo que o modelo proposto, a menos do vocabulário, mostra-se abrangente o suficiente para que sua extensão ocorra com custo mínimo.

Referências

Bakhtin, M. ([1953]) "Gêneros do discurso", In A estética da criação verbal, Martins Fontes, São Paulo, 2000.

Dorr, B., Jordan, P., and Benoit, J. (1999) "A Survey of Current Research in Machine Translation," Advances in Computers, Vol 49, M. Zelkowitz (Ed), Academic Press, London, pp. 1-68.

Felipe, T. A. F. S. (1998) *A Relação Sintático-Semântica dos Verbos e seus Argumentos na Língua Brasileira de Sinais*. Tese de doutorado, UFRJ.

Hutchins, W. J. and Somers, H. L. (1992) An introduction to Machine Translation, Academic Press, San Diego (CA).

Uchida. H., Zhu, M and Della Senta, T. (1999) A *gift for a millenium,* IAS/UNU, Tokyo.

Wojcik, R. and Hoard, J. (1995) "Controlled languages in industry", In: Survey of the State of the Art in Human Language Technology, Edited by Cole, R.A.; Mariani, J.; Uszkoreit, H.; Zaenen, A.; Zue, V., http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html, November.

Zimmermann, G. and Vanderheiden, Gregg (2001) Translation on Demand Anytime and Anywhere. CSUN's 16th International Conference, March 19 - 24, 2001, Los Angeles, CA

Santaella, L. (2001). Matrizes da linguagem e pensamento, Iluminuras, São Paulo.