Universidade de São Paulo - USP

Universidade Federal de São Carlos - UFSCar

Universidade Estadual Paulista - UNESP

A construção do corpus e dos dicionários Inglês-UNL e UNL-português para o projeto EPT-Web

Lucas Antiqueira

Marcela Franco Fossey

Tatiana Pedrolongo

Juliana Galvani Greghi

Ronaldo Teixeira Martins

Maria das Graças Volpe Nunes

NILC-TR-02-24

Dezembro, 2002

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Este relatório apresenta o processo de construção dos corpora e dos dicionários Inglês-UNL e UNL-Português para o projeto EPT-Web, que visa a tradução automática para o português de textos jornalísticos (manchetes e leads) produzidos em língua inglesa e disponíveis na versão eletrônica da primeira página do jornal norte-americano The New York Times. O projeto é realizado pelo NILC e tem o apoio do CNPq.

Índice

1 Introdução

O problema em torno do qual se organiza o projeto EPT-Web[1] é o desenvolvimento de uma ferramenta computacional para tradução para a língua portuguesa (ou seja, unidirecional), de forma completamente automática, de textos jornalísticos (manchetes e leads) produzidos em língua inglesa e disponíveis na versão eletrônica da primeira página de um diário norte-americano específico (The New York Times). Trata-se, em última instância, do desenvolvimento de um plug-in a ser incorporado a uma ferramenta de navegação na Web (Internet Explorer), que permita, ao usuário, a tradução, em tempo real, e com qualidade, da página inicial da versão eletrônica do jornal americano (http://www.nyt.com), preservada a sua diagramação, figuras, tabelas, links e outros recursos. O aplicativo representa uma tentativa de aproximação experimental, gradual e controlada ao problema mais geral - e cuja solução se estende por prazo muito mais longo - do desenvolvimento de um sistema de tradução automática de natureza mais genérica e de alcance mais amplo, não restrito quanto à fonte (The New York Times), quanto ao domínio (noticiário geral), quanto ao gênero (informativo de natureza jornalística) e mesmo quanto à língua (inglês) dos textos a serem submetidos à tradução. A execução do projeto adota um modelo de tradução multilingual baseado em interlíngua, derivado do Projeto UNL.

A escolha da UNL como instrumento de tradução tem também motivação empírica. O Projeto inclui hoje 15 línguas, cada uma das quais atribuída a diferentes grupos de pesquisa e desenvolvimento espalhados pelo mundo: árabe, alemão, chinês, espanhol, francês, hindi, indonésio, inglês, italiano, japonês, letão, mongol, português, russo e tailandês. Tem duração prevista de dez anos, a partir de 1997, e seu objetivo final é o desenvolvimento de um sistema que habilite cada usuário da Internet a produzir material em sua própria língua materna e disponibilizá-lo em UNL, por meio do codificador língua natural-UNL, para que outros usuários, falantes de outras línguas, possam ter acesso a esse material, empregando os decodificadores UNL-língua natural. Os sistemas de codificação e decodificação estão sendo desenvolvidos por universidades, institutos de pesquisa e empresas de diversos países, o que distingue o projeto UNL como uma tentativa de conjugar esforços de especialistas em processamento automático das línguas naturais (PLN) do mundo todo. As ferramentas de PLN para o português do Brasil (UNL-Brasil), estão a cargo do Núcleo Interinstitucional de Lingüística Computacional - NILC [2].

A produção de sistemas de tradução automática exige, hoje, na maioria das vezes, intervenção humana, na pré-edição do texto de entrada, e/ou pós-edição do texto final. Isso ocorre por causa dos resultados produzidos que ainda são pífios [1], o que vem conduzindo, mais recentemente [2], a uma revisão de metas, segundo a qual a tradução completamente automática, com qualidade, de textos arbitrários, não será atingida em curto prazo. Para a produção de resultados aceitáveis, este projeto circunscreve o seu campo de ação imediato a um gênero textual bastante determinado (manchetes e leads de textos jornalísticos), de dimensões reduzidas, estruturas fixas e conteúdos literais. Para salvaguardar a portabilidade

[1] Financiado pelo CNPq (# 551485/2001-9) http:// www.nilc.icmc.usp.br/nilc/projects/ept-web.htm
[2] http://www.nilc.icmc.sc.usp.br

do sistema, definiu-se, porém, um domínio (o noticiário) menos específico que pudesse permitir o tratamento das ambigüidades normalmente encontradas em textos mais gerais. No entanto, para que as estratégias e recursos desenvolvidos não fiquem restritos ao funcionamento de um sistema de tradução automática excessivamente especializado, é utilizada, como instrumento metodológico, um subconjunto da Universal Networking Language (UNL), linguagem artificial para representação semântica do conteúdo ideacional dos enunciados lingüísticos, desenvolvida pelo Instituto de Estudos Avançados da United Nations University, em Tóquio, e posta em domínio público em 1999 [3,4]. A utilização da UNL como interlíngua semântica entre o inglês e português assegura ao sistema a portabilidade necessária para sua expansão para outros domínios e gêneros (em função do caráter abrangente e generalista da UNL) e para uma plataforma multilíngüe (dado que a incorporação de novas línguas aos sistemas interlinguais pode ser feita de forma independente e modular, com a cooperação de outros grupos, também envolvidos com o desenvolvimento de sistemas para UNL).

Nesse contexto, este relatório apresenta a descrição das atividades realizadas durante os primeiros 12 dos 24 meses previstos de projeto, mais especificamente aquelas que envolvem a construção dos corpora e dos dicionários inglês-UNL e UNL-português. A Seção 2 descreve a metodologia utilizada para a criação do corpus de trabalho e de teste; a Seção 3 discute a criação dos dicionários Inglês-UNL e UNL-Português; a Seção 4 indica os próximos passos do projeto, e a Seção 5 conclui este relatório.

2 Metodologia de criação do corpus

A compilação de um corpus em projetos como esse tem o objetivo de realizar um levantamento do conhecimento necessário para o desenvolvimento do sistema e caracterizar profunda e detalhadamente, para efeito de tradução automática, as características fundamentais do gênero de texto investigado. Para tanto, o conjunto de textos coletado deve ser processado para que haja a diferenciação entre textos a serem submetidos ao tradutor e marcações de layout, que devem ser reproduzidas na reconstrução da página traduzida. O conjunto de textos deve, depois de processado, ser analisado para que sejam definidos: uma subespecificação da UNL adequada ao gênero jornalístico, e os dicionários que devem ser usados para o desenvolvimento da aplicação. Para tala tarefa, foi usado um subconjunto de representativo de textos (corpus), extraído da versão eletrônica do jornal "The New York Times". Nas seções seguintes serão apresentados o processo de obtenção das páginas e o tratamento que as mesmas receberam.

2.1. Construção do corpus

As primeiras páginas do jornal "The New York Times" [Figura 1] foram coletadas durante aproximadamente um mês, a cada meia hora, com o auxílio da ferramenta Wget[3]. Essa ferramenta permite que a coleta de páginas possa ser realizada sem supervisão humana, ou seja, não é necessária a presença de um usuário para que o processo de cópia das páginas possa ocorrer. Uma particularidade desse corpus é que as páginas foram coletadas com

[3] GNU Wget Manual, http://www.gnu.org/manual/wget/, disponível em 25 de agosto de 2002.

intervalos de 30 minutos. A ferramenta Wget oferece, como uma de suas funcionalidades, a opção de salvar o arquivo indicado pela URL em questão periodicamente, bastando, para isso, indicar o intervalo (em segundos) que deverá ser usado entre coletas sucessivas.
Para que tal tarefa pudesse ser realizada adequadamente, foi necessário desenvolver um pequeno aplicativo em linguagem C que renomeasse os arquivos salvos antes que os mesmos fossem sobrescritos na próxima coleta. Os arquivos foram renomeados seqüencialmente, seguindo o seguinte padrão: index.html.<numeração seqüencial>

Figura 1 – Exemplos de primeira página do jornal "The New York Times"

O corpus foi dividido em dois subconjuntos: o de trabalho, construído a partir das páginas coletadas a cada seis horas, e o de teste, baseado nas páginas restantes. O corpus de trabalho é o que será usado na construção dos dicionários e regras/gramática de tradução que apoiarão a tarefa de tradução automática; e o corpus de teste será usado na validação do protótipo.

2.2. Pós-processamento do Corpus

Os arquivos resultantes da coleta do corpus necessitam de pós-processamento para que possam ser utilizados para o desenvolvimento da aplicação. Esses arquivos contêm "informações" irrelevantes, que podem interferir no trabalho com o corpus. Essas informações são *tags* HTML, utilizadas para fazer a formatação da página tal como ela é apresentada ao usuário, que poderiam interferir no trabalho com o corpus já que são usadas palavras em inglês para marcar as informações e elas poderiam ser confundidas com o que denominamos neste trabalho "unidades sentenciais". As unidades sentenciais são entendidas aqui como sendo todas as construções analisáveis que aparecem nos arquivos do

corpus, por exemplo, os menus de opções e as indexações para seções especiais como "negócios", "previsão do tempo" ou "ciência". Além das *tags* foi necessário realizar uma seleção das unidades sentenciais com o objetivo de eliminar sentenças repetidas que será reportado posteriormente.

2.2.1. Eliminação de tags HTML

Foi desenvolvido, paralelamente ao processo de coleta, um filtro automático com a função de retirar todo o texto que não deveria ser levado em consideração na tradução do Inglês para o Português (código HTML e Java, por exemplo). Esse filtro poderia ser construído com uma linguagem de programação tradicional ou com uma ferramenta cuja utilização fosse mais orientada à solução do problema em questão.

Para que tal decisão fosse tomada, foram realizados testes com a linguagem C e com a ferramenta Flex[4]. Notou-se que a simples retirada do código HTML (caracteres delimitados por parênteses angulares, < e >) resultava numa saída ainda com elementos indesejáveis, como códigos em Java Script e várias linhas vazias (em geral, em um arquivo HTML, as quebras de linha não estão dentro de um *tag* e, por isso, permaneceram no texto mesmo depois da utilização do filtro).

Outro problema é que a linguagem HTML permite, também, a inserção de códigos especiais fora da delimitação dos *tags*, usados para a exibição de caracteres acentuados ou espaços em branco. Nas páginas coletadas encontramos apenas o código para espaços em branco (" ").

A ferramenta Flex se mostrou mais adequada para a solução do problema, já que para sua utilização é necessário, apenas, que seja construído um código onde são definidas regras formadas por padrões a serem identificados num texto de entrada e ações, que refletirão diretamente no texto de saída, a serem executadas no caso de tais padrões serem identificados. Este código é processado pela ferramenta e um código em linguagem C é gerado.

Em linhas gerais, o filtro Flex construído elimina todos os elementos indesejáveis discutidos anteriormente, dispondo o texto restante em sentenças em língua natural. O conjunto de tags consecutivos, que não contém texto em língua natural, é transformado em uma quebra de linha na saída formatada.

O filtro gerado recebe o código HTML da entrada padrão e envia seus resultados para a saída padrão. Para executá-lo com arquivos, basta utilizar o recurso de redirecionamento de entrada e saída padrões, que alguns sistemas operacionais fornecem (UNIX e DOS/Windows têm essa característica). Também foi construído um arquivo em lote do DOS/Windows (as páginas coletadas foram pós-processadas nesta plataforma) que envia para o filtro, através do redirecionamento, todos os arquivos HTML contidos num determinado diretório, e coloca os arquivos filtrados num diretório separado dos arquivos originais.

2.2.2. Eliminação de sentenças repetidas

[4] Flex - A fast scanner generator, http://www.gnu.org/manual/flex-2.5.4/flex.html, disponível em 25 de agosto de 2002.

A última tarefa de pós-processamento realizado foi a eliminação de unidades sentenciais repetidas do conjunto de páginas pós-processadas. Como as páginas foram capturadas a cada 30 minutos, várias unidades presentes em um arquivo capturado são encontradas, também, nos arquivos subseqüentes, e para evitar que as unidades fossem consideradas e também analisadas mais de uma vez, foram investigados métodos/soluções já existentes, que tratassem problemas semelhantes. Foi encontrada uma solução relacionada indiretamente ao problema: uma ferramenta eficiente que realiza a junção (merge)[5] de múltiplos arquivos ordenados linha a linha, com as sentenças repetidas dispostas consecutivamente. Foi desenvolvido, em Perl, um programa que elimina as linhas repetidas, computando o número de repetições para cada linha diferente.

Neste ponto, quando a coleta das páginas iniciais do jornal já havia terminado e as ferramentas para a manipulação destas já estavam disponíveis, o corpus de trabalho pôde ser construído. O filtro foi aplicado no conjunto de páginas selecionadas e foram obtidos arquivos pós-processados num diretório à parte; já em ambiente FreeBSD, os arquivos foram, um a um, convertidos e ordenados, respectivamente, pelos utilitários Dosunix e Sort. Os arquivos foram submetidos aos pares (um dos arquivos usados em um par foi o resultado da união de pares anteriores) à ferramenta de junção e, finalmente, as sentenças repetidas foram eliminadas quase totalmente. Podem ser encontradas, ainda, ocorrências como o exemplo abaixo:

```
Anthony Zinni met with Yasir Arafat today as Israel's offensive against
Palestinian militants entered its second week.

Anthony Zinni met with Yasir Arafat today, as Israel's offensive against
Palestinian militants entered its second week.
```

Neste caso, a ferramenta identifica duas sentenças diferentes já que na segunda ocorrência há uma “ , “ depois de *today*. Essas ocorrências não puderam ser eliminadas automaticamente, mas tais ocorrências são escassas.

2.3. Estatísticas do Corpus

No processo de cópia das páginas do jornal, durante o período de 3 abril a 8 de maio de 2002, obtivemos 1.462 páginas, coletadas a cada meia hora, das quais 150 formam o corpus de trabalho. Estas páginas do corpus de trabalho foram coletadas numa freqüência de seis horas. O restante das páginas forma o corpus de teste. O processamento das páginas coletadas, descrito anteriormente, foi aplicado na formação do corpus de trabalho, que contém 47.691 palavras (não necessariamente distintas) e 5.291 sentenças.

Não foram encontradas dificuldades nesta etapa inicial do projeto. Foi necessária, apenas, a familiarização com a ferramenta Flex, em substituição à linguagem C, na construção do filtro para as páginas coletadas.

[5] multimerge.inc - Dan Melamed's NLP Research Software Library - General Processing Section http://www.cs.nyu.edu/~melamed/genproc.html, disponível em 25 de agosto de 2002.

3 Criação dos dicionários Inglês-UNL e UNL-Português

Para a concretização do projeto, foi necessário, primeiramente, caracterizar profunda e detalhadamente, para efeito de tradução automática, o gênero de texto sob investigação: manchetes e leads das versões eletrônicas das primeiras páginas dos jornais americanos. Foi também necessário subespecificar o modelo de representação interlingual adotado (UNL), definindo o conjunto de entidades semânticas mais adequado à representação dos conteúdos veiculados pelo gênero abordado.

A caracterização foi uma etapa importante devido à necessidade de levantamento do vocabulário e das estruturas sintáticas empregadas neste gênero de texto na língua inglesa e na língua portuguesa. O vocabulário vai determinar as entradas do dicionário e as estruturas sintáticas, o mapeamento. Este estudo detalhado do gênero jornalístico possibilitou também a subespecificação da UNL. Por meio dele, pôde-se eliminar recursos da UNL desnecessários e criar novos mais adequados para este trabalho.

A construção efetiva dos dicionários após a realização dessas análises foi iniciada.

3.1. Elaboração do Dicionário

A seguir são descritas as atividades realizadas, até o momento, relativas ao desenvolvimento dos recursos lingüísticos necessários para a elaboração do dicionário.

3.1.1. Números

O total de páginas capturadas foi de 1462 páginas, com uma média de 250 sentenças por página (total aproximado de 23500 sentenças distintas). O corpus total foi dividido em corpus de trabalho e corpus de teste. Para a elaboração do dicionário, foi utilizado somente o corpus de trabalho, que foi dividido em três arquivos: corpus1 (30 páginas), corpus2 (59 páginas) e corpus3 (58 páginas).

Até o momento, trabalhamos apenas com o corpus1, o que resultou num número total de 8901 entradas (3754 palavras e 5147 sufixos verbais).

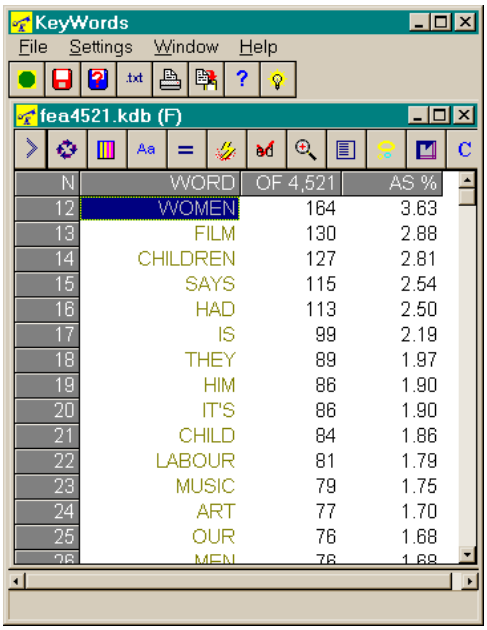
3.1.2. Organização

O dicionário Inglês – UNL – Português foi construído no Microsoft Access. Trabalhamos com os dois dicionários ao mesmo tempo e, depois de prontos, serão divididos em dois: Inglês – UNL / UNL – Português. A primeira versão é dividida em 10 campos, sendo eles: **UW** (Universal Word), **NormalEnglish** (entrada da palavra no dicionário em inglês), **English** (variações de gênero e número, quando tais variações não podem ser geradas automaticamente, e conjugações verbais), **NormalPortuguese** (entrada da palavra no dicionário em português), **Portuguese** (variações de gênero e número, quando tais variações não podem ser geradas automaticamente, ou a raiz de verbos e de palavras que variam em número e gênero, mas seguem um padrão sintático), **FeaturesEnglish** (características morfológicas de cada entrada das palavras em inglês), **FeaturesPortuguese** (características morfológicas de cada entrada das palavras em português), **FrequenceEnglish** (indica a freqüência de ocorrência da palavra no inglês, funciona como um critério de desambigüização pela freqüência de ocorrência e só é preenchido quando

temos competição entre duas entradas), **FrequencePortuguese** (indica a freqüência de ocorrência da palavra no português, funciona como um critério de desambigüização pela freqüência de ocorrência e só é preenchido quando temos competição entre duas entradas), e **ID** (é apenas um índice para organização interna do dicionário e é preenchido automaticamente cada vez que o dicionário é compactado para agilizar o processo de armazenagem e busca).

3.1.3. Elaboração

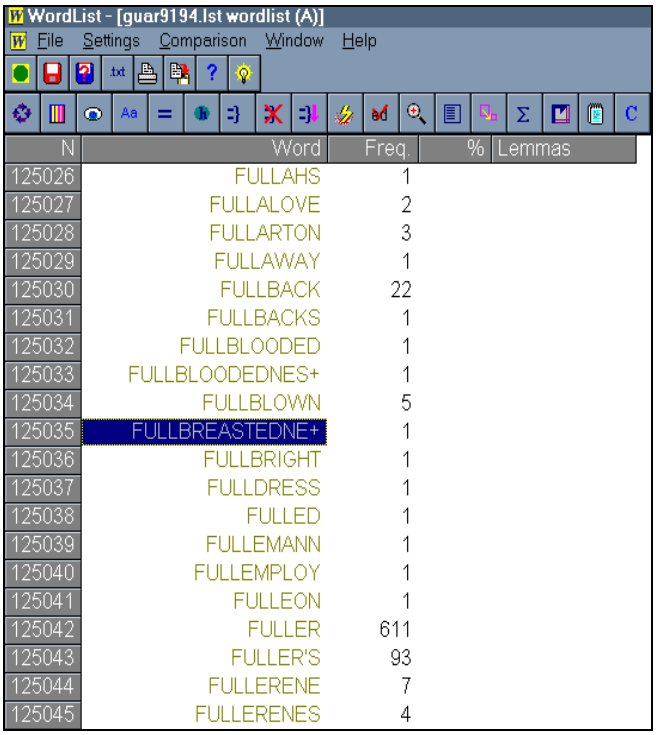
Foram gerados três arquivos com sentenças não repetidas, para codificação Inglês-UNL e decodificação UNL-Português. Esses dicionários estão sendo construídos com o auxílio da ferramenta WordSmith[6], desenvolvida na Universidade de Oxford. Esta é uma ferramenta para análise lexical que permite que sejam realizadas várias operações, entre elas: listar as palavras presentes em um texto, gerar dados estatísticos como o número total de palavras de um corpus, o tamanho das palavras e o número total de sentenças, analisar palavras em um determinado contexto e identificar palavras-chave. Para evitar a sobrecarga dessa ferramenta, o corpus foi divido em três arquivos que serão, posteriormente, agrupados em um único arquivo. As figuras 2 e 3 ilustram algumas das funcionalidades do WordSmith.



N	WORD	OF 4,521	AS %
12	WOMEN	164	3.63
13	FILM	130	2.88
14	CHILDREN	127	2.81
15	SAYS	115	2.54
16	HAD	113	2.50
17	IS	99	2.19
18	THEY	89	1.97
19	HIM	86	1.90
20	IT'S	86	1.90
21	CHILD	84	1.86
22	LABOUR	81	1.79
23	MUSIC	79	1.75
24	ART	77	1.70
25	OUR	76	1.68

Figura 2 – Lista de palavras-chave

[6] Wordsmith Tools, http://www.liv.ac.uk/~ms2928/index.htm, disponível em 25 de agosto de 2002.



N	Word	Freq.	%	Lemmas
125026	FULLAHS	1		
125027	FULLALOVE	2		
125028	FULLARTON	3		
125029	FULLAWAY	1		
125030	FULLBACK	22		
125031	FULLBACKS	1		
125032	FULLBLOODED	1		
125033	FULLBLOODEDNES+	1		
125034	FULLBLOWN	5		
125035	FULLBREASTEDNE+	1		
125036	FULLBRIGHT	1		
125037	FULLDRESS	1		
125038	FULLED	1		
125039	FULLEMANN	1		
125040	FULLEMPLOY	1		
125041	FULLEON	1		
125042	FULLER	611		
125043	FULLER'S	93		
125044	FULLERENE	7		
125045	FULLERENES	4		

Figura 3 – Lista de palavras extraídas de um texto

As acepções de cada palavra foram escolhidas de acordo com o contexto do corpus1. Ao mesmo tempo em que essa atitude tornou o trabalho mais simples, uma vez que colocar todas as acepções possíveis de uma palavra tornaria a construção do dicionário muito extensa, ela delimitou bastante a abrangência do dicionário, uma vez que ele só contém as acepções do corpus em questão.

Para o preenchimento de todos os campos do dicionário (UW, NormalEnglish, NormalPortuguese,...), tanto o comportamento morfológico quanto o comportamento sintático de cada palavra foi levado em consideração. Primeiro, foi avaliado se a palavra era verbo, substantivo, adjetivo, advérbio, conjunção, preposição ou pronome. Depois de decidido a qual classe morfológica a palavra pertencia e qual a acepção mais adequada, era observado o seu comportamento na sentença: para os adjetivos, era pertinente saber se era um adjetivo predicativo ou atributivo. No caso dos verbos, se eram verbos de processo, ação ou estado. Para cada uma dessas situações, existe uma restrição, que é colocada na UW, sendo (aoj>thing), para adjetivos atributivos; (mod<thing) para adjetivos predicativos; (icl>be) para verbos de estado; (icl>do) para verbos de ação e (icl>occur) para verbos de processo. Essas diferenciações podem ser observadas na Tabela 1:

Tabela 1 – Organização do dicionário Inglês – UNL – Português.

"UW"	Normal English	English	Normal Portuguese	Portuguese	EnglishFeatures	PortugueseFeatures	Frequence Portuguese	Frequence English	ID
"believe(icl>be)"	believe	[believe]	acreditar	[acredit]	(VER,VD2,RT1,STA)	(VER,P05,STM,VI2,!em,STA)	<P,0,0>		{}
"believe(icl>be)"	believe	[believed]	acreditar	[acredit]	(VER,VD2,RT0,PTP,STA)	(VER,P05,STM,VI2,!em,STA)	<P,0,0>		{}
"believe(icl>be)"	believe	[believes]	acreditar	[acredit]	(VER,VD2,RT1,3PS,STA)	(VER,P05,STM,VI2,!em,STA)	<P,0,0>		{}
"dawn(icl>occur)"	dawn	[dawns]	amanhecer	[amanhe]	(VER,VD1,RT1,3PS,PRO)	(VER,VD1,P16,STM,PRO)	<P,0,0>		{}
"dawn(icl>occur)"	dawn	[dawning]	amanhecer	[amanhe]	(VER,VD1,GER,PRO)	(VER,VD1,P16,STM,PRO)	<P,0,0>		{}
"dawn(icl>occur)"	dawn	[dawn]	amanhecer	[amanhe]	(VER,VD1,RT1,PRO)	(VER,VD1,P16,STM,PRO)	<P,0,0>		{}
"dawn(icl>occur)"	dawn	[dawned]	amanhecer	[amanhe]	(VER,VD1,RT0,PTP,PRO)	(VER,VD1,P16,STM,PRO)	<P,0,0>		{}
"respond(icl>do)"	respond	[respond]	responder	[respond]	(VER,VD2,RT1,ACT)	(VER,P06,STM,VD2,ACT)	<P,0,0>		{}
"respond(icl>do)"	respond	[responds]	responder	[respond]	(VER,VD2,RT1,3PS,ACT)	(VER,P06,STM,VD2,ACT)	<P,0,0>		{}
"respond(icl>do)"	respond	[responded]	responder	[respond]	(VER,VD2,RT0,PTP,ACT)	(VER,P06,STM,VD2,ACT)	<P,0,0>		{}
"respond(icl>do)"	respond	[responding]	responder	[respond]	(VER,VD2,GER,ACT)	(VER,P06,STM,VD2,ACT)	<P,0,0>		{}
"respond(icl>do)"	respond	[respond]	responder	[respond]	(VER,VD2,RT1,ACT)	(VER,P06,STM,VD2,ACT)	<P,0,0>		{}
"brief(mod<thing)"	brief	[brief]	breve	[breve]	(ADJ)	(ADJ,COM)	<P,0,0>		{}
"deadly(mod<thing)"	deadly	[deadly]	letal	[leta]	(ADJ)	(ADJ,COM,STM,SGN.l,PLN.is)	<P,0,0>		{}
"false(aoj>thing)"	false	[false]	falso	[fals]	(ADJ)	(ADJ,STM)	<P,0,0>		{}
"high(aoj>thing)"	high	[high]	elevado	[elevad]	(ADJ)	(ADJ,STM)	<P,0,0>		{}

3.1.4 Dificuldades Encontradas

Os problemas encontrados dizem respeito à diferenciação de palavras homônimas. No caso de palavras homônimas com classes morfológicas distintas, o problema, como podemos observar na Tabela 2, foi facilmente resolvido, uma vez que existem restrições para cada uma das classes. Vale lembrar que essas restrições foram usadas apenas em caso de ambigüidade dentro do conjunto de palavras do dicionário.

Tabela 2 – Restrições para palavras homônimas com classes morfológicas distintas.

"UW"	Normal English	English	Normal Portuguese	Portuguese	EnglishFeatures	PortugueseFeatures	Frequence Portuguese	Frequence English	ID
“back(icl>do)”	back	[back]	amparar	[ampar]	(VER,VD2,RT1,ACT)	(VER,P05,STM,VD2,ACT)	<P,0,0>		{}
“back(icl>do)”	back	[backs]	amparar	[ampar]	(VER,VD2,RT1,3PS,ACT)	(VER,P05,STM,VD2,ACT)	<P,0,0>		{}
“back(icl>do)”	back	[backed]	amparar	[ampar]	(VER,VD2,RT0,PTP,ACT)	(VER,P05,STM,VD2,ACT)	<P,0,0>		{}
“back(icl>do)”	back	[backing]	amparar	[ampar]	(VER,VD2,GER,ACT)	(VER,P05,STM,VD2,ACT)	<P,0,0>		{}
“back(icl>thing)”	back	[back]	costas	[costas]	(NOU)	(NOU,PLN)	<P,0,0>		{}
“back(icl>how)”	back	[back]	de_volta	[de volta]	(ADV)	(ADV)	<P,0,0>		{}

Já no caso de palavras homônimas de mesma classe morfológica, a solução não foi tão facilmente encontrada. Nesses casos, recorremos à análise semântica, e criamos algumas restrições, que representam sinônimos dessas palavras, para diferenciá-las, conforme observa-se na Tabela 3.

Tabela 3 – Restrições para palavras homônimas com classes morfológicas distintas.

	Normal English	English	Normal Portuguese	Portuguese	EnglishFeatures	PortugueseFeatures	Freq. Port.	Freq. Eng.	ID
"age(icl>period)"	age	[age]	época	[época]	(NOU)	(NOU,FEM)	<P,0,0>		{}
"age(icl>time)"	age	[age]	idade	[idade]	(NOU)	(NOU,FEM)	<P,0,0>		{}
"reason(icl>motive)"	reason	[reason]	razão	[raz]	(NOU)	(NOU,FEM,STM,SGN.ão,PLN.ões)	<P,0,0>		{}
"reason(icl>thinking)"	reason	[reason]	razão	[razão]	(NOU)	(NOU,FEM,SGN)	<P,0,0>		{}
"ask(icl>do(agt>human,obj>question))"	ask	[ask]	perguntar	[pergunt]	(VER,VD3,RT1,ACT)	(VER,P05,STM,VD2,ACT)	<P,0,0>		{}
"ask(icl>do(agt>human,obj>question))"	ask	[asks]	perguntar	[pergunt]	(VER,VD3,RT1,3PS,ACT)	(VER,P05,STM,VD2,ACT)	<P,0,0>		{}
"ask(icl>do(agt>human,obj>question))"	ask	[asked]	perguntar	[pergunt]	(VER,VD3,RT0,PTP,ACT)	(VER,P05,STM,VD2,ACT)	<P,0,0>		{}
"ask(icl>do(agt>human,obj>question))"	ask	[asking]	perguntar	[pergunt]	(VER,VD3,GER,ACT)	(VER,P05,STM,VD2,ACT)	<P,0,0>		{}
"ask(icl>do(agt>human,obj>thing))"	ask	[ask]	pedir	[pe]	(VER,VD3,RT1,ACT)	(VER,P51,STM,VD2,ACT)	<P,0,0>		{}
"ask(icl>do(agt>human,obj>thing))"	ask	[asks]	pedir	[pe]	(VER,VD3,RT1,3PS,ACT)	(VER,P51,STM,VD2,ACT)	<P,0,0>		{}
"ask(icl>do(agt>human,obj>thing))"	ask	[asked]	pedir	[pe]	(VER,VD3,RT0,PTP,ACT)	(VER,P51,STM,VD2,ACT)	<P,0,0>		{}
"ask(icl>do(agt>human,obj>thing))"	ask	[asking]	pedir	[pe]	(VER,VD3,GER,ACT)	(VER,P51,STM,VD2,ACT)	<P,0,0>		{}

No entanto, algumas vezes o problema de homonímia não pôde ser solucionado. Foram casos em que além de homonímia, temos também a sinonímia entre as palavras. Nem a análise sintática nem a semântica são suficientes para possibilitar a diferenciação. Nesses casos, há uma questão relacionada ao estilo que envolve as sutilizas de cada língua e que a UNL ainda não possibilita representar. Alguns desses casos estão representados na Tabela 4.

Tabela 4 – Problemas de homonímia e sinonímia: sem solução.

"UW"	Normal English	English	Normal Portuguese	Portuguese	EnglishFeatures	PortugueseFeatures	Frequence Portuguese	Frequence English	ID
"face(icl>do)"	face	[face]	encarar	[encar]	(VER,VD2,RT1,ACT)	(VER,P05,STM,VD2,ACT)	<P,0,0>		{}
			defrontar				<P,0,0>		{}
"beautiful(aoj>thing)"	beautiful	[beautiful]	bonito	[bonit]	(ADJ)	(ADJ,STM)	<P,0,0>		{}
			belo				<P,0,0>		{}
"bald(aoj>thing)"	Bald	[bald]	calvo	[calv]	(ADJ)	(ADJ,STM)	<P,0,0>		{}
			careca				<P,0,0>		{}

Dentro das possibilidades que nos estão disponíveis, a única solução foi escolher entre uma das opções, o que acaba por diminuir a variedade de vocabulário, empobrecendo o dicionário.

4 Conclusão e Próximos passos

As tarefas realizadas até o momento se mostraram relevantes para a continuidade do projeto e foram desenvolvidas de maneira satisfatória, no tempo previsto.

O que foi feito até agora é a primeira parte de um projeto mais extenso, que tem por objetivo final a tradução totalmente automática das primeiras páginas da versão digital do New York Times, como já mencionado anteriormente. Os dicionários que estão prontos foram feitos com base unicamente no corpus1. No momento, está em andamento a codificação manual Inglês-UNL deste mesmo corpus. Terminada esta segunda fase,

daremos início à complementação do dicionário, utilizando os corpora 2 e 3, para finalmente codificar manualmente esses corpora, seguindo o mesmo procedimento utilizado para o corpus1.

Vale lembrar que concomitantemente com essa parte lingüística do desenvolvimento do projeto, estão sendo desenvolvidas a ferramenta de codificação inglês-UNL para o mapeamento das estruturas do texto fonte na representação interlingual adotada, e a ferramenta de decodificação UNL-português para o mapeamento das estruturas da UNL em sentenças gramaticais da língua portuguesa.

5 Referências Bibliográficas

[1] Oliveira Jr. O.N.; Martins, M.S.; Marchi, A.R.; Martins, R.T. (2000) A critical analysis of the performance of English-Portuguese-English MT systems. Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'2000), pp.85-92. Atibaia, SP.

[2] Cole, R.A.; Mariani, J.; Uszkoreit, H.; Zaenen, A.; Zue, V. (eds.) (1995). Survey of the State of the Art in Human Language Technology. NSF/CEC/CSLU. Oregon Graduate Institute. November. (http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html)

[3] UNL (1996). UNL: Universal Networking Language - An Electronic Language for Communication, Understanding and Collaboration. UNU/IAS/UNL Center. Tokyo, Japan.

[4] Uchida, H.; Zhu, M.; Della Senta, T. (1999). The UNL, a Gift for a Millennium. UNU/IAS, 237, November.

6 Bibliografia Consultada

MARTINS, R.T. et al. O uso de interlíngua para comunicação via Internet: O Projeto UNL/Brasil. Série de RELATÓRIOS do NILC. São Carlos: [s.n], 2001 (NILC-TR-01-3).

MARTINS, R.T.; NUNES, M.G.V.; RINO, L.H.M. As Regras Gramaticais para a Decodificação UNL-Português no Projeto UNL. Série de Relatórios do NILC, NILC-TR-98-1.

Nunes M.G.V.; Rino L.H.M.; Sossolote C.R.C.; Zavaglia C. (1997). As Manifestações Morfossintáticas da Linguagem UNL no Português do Brasil. Série de Relatórios do NILC, NILC-97-TR-2.

Ryan, M.A.F. (1993). Conjugação dos Verbos em Português. São Paulo: Ática.

Swan, M. (1995). Practical English Usage. Oxford: Oxford University Press, pp.