UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

O Uso de Interlíngua para Comunicação via Internet: O Projeto UNL/Brasil

Osvaldo Novais de Oliveira Jr.
Ronaldo Teixeira Martins
Lúcia Helena Machado Rino
Maria das Graças Volpe Nunes

Nº 61

NOTAS

São Carlos - SP

Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

O uso de interlíngua para comunicação via Internet: O Projeto UNL/Brasil

Osvaldo Novais de Oliveira Jr.
Ronaldo Teixeira Martins
Lucia Helena Machado Rino
Maria das Graças Volpe Nunes

NILC-TR-01-3

Julho, 2001

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

O uso de interlíngua para comunicação via Internet: O Projeto UNL/Brasil

Maria das Graças Volpe Nunes
ICMC-USP
mdgvnune@icmc.sc.usp.br

Lúcia H. Machado Rino
DC-UFSCar
lucia@dc.ufscar.br

Ronaldo Teixeira Martins
CCHS-USF
ronaldo@nilc.icmc.sc.usp.br

Osvaldo N. Oliveira Jr.
IFSC-USP
chu@ifsc.sc.usp.br

NILC-São Carlos
ICMC-USP
C.P. 669
13560-970 São Carlos, SP

RESUMO

A despeito da disponibilidade da tecnologia da Internet, um dos principais problemas enfrentados por seus encontra-se na barreira da língua: o inglês, na maioria das vezes, é a língua franca nesse meio, embora apenas 8% da população mundial tenham nela sua língua materna. Para diminuir essa barreira, o projeto descrito a seguir apresenta uma proposta de interlíngua para incorporar a capacidade de tradução automática de textos na WWW, por intermédio de ferramentas de *software* para codificação e decodificação de textos em línguas naturais. Essas ferramentas serão acopladas aos *browsers* disponíveis, de maneira a permitir que usuários produzam e leiam textos em sua própria língua. Este artigo apresenta uma breve descrição da interlíngua denominada *Universal Networking Language*, ou UNL, assim como as metodologias e resultados alcançados para a decodificação do português. As limitações da linguagem UNL, assim como possíveis aplicações, são também discutidas.

ABSTRACT

In the past few years, there has been an explosive growth of the World Wide Web (WWW), which has become the most powerful and useful means to access information in the Internet, for retrieval purposes. In spite of the availability of such a technology, the major problem faced by Internet users lies on the language barrier: the English language, most often, is the *lingua franca* in the Internet, although only 8% of the world population have English as their native language. To alleviate this barrier, the project presented here describes an interlingua proposal to incorporate the capability of language translation into the WWW. Software tools that allow for codification and decoding of texts have been developed and shall be encapsulated in available browsers, so that users throughout the world can write and read texts in their own language. This article presents a brief description of the interlingua called *Universal Networking Language*, or UNL, and the methodologies and results already obtained in decoding UNL into Portuguese. The limitations of the UNL language, as well as some of its potential applications, are also discussed.

Palavras-Chaves: interlíngua, tradução automática, comunicação via Internet

O uso de interlíngua para comunicação via Internet: O Projeto UNL/Brasil

1. Introdução

A Internet é uma promessa de acesso fácil e barato à informação que, em princípio, tende a ser democratizante, no sentido de que populações de menor poder aquisitivo possam se beneficiar desse recurso da mesma forma que os extratos mais privilegiados da sociedade. Essa promessa provavelmente se tornará realidade em breve, pelo menos do ponto de vista da tecnologia das comunicações. Assim como já acontece com a TV e o vídeo cassete, é provável que num futuro bastante próximo a Internet esteja disponível na maioria dos lares, mesmo em países em desenvolvimento como o Brasil. Infelizmente esse cenário não é tão promissor quanto parece. A despeito da disponibilidade da tecnologia, o acesso à informação requer uma preparação dos usuários, só atingível com programas de educação de médio e longo prazos. Neste contexto, uma das maiores barreiras é a da língua. Muito do que aparece na Internet está em inglês porque foi produzido em países de língua inglesa, como os Estados Unidos, ou porque o produtor desejava atingir uma audiência mais abrangente. De fato, o inglês passou a ser a língua franca da Internet, embora apenas 8% da população mundial tenham no inglês sua língua materna.

Talvez ainda não percebamos a extensão das limitações impostas com o uso do inglês para comunicação na Internet. Afinal, atualmente os usuários da Internet no Brasil pertencem às classes mais favorecidas, em que a probabilidade de se ter um conhecimento da língua inglesa é relativamente alta. À medida que a Internet se popularizar, entretanto, a barreira da língua se fará sentir com grande força, mesmo porque não é imaginável que grandes massas da população aprendam inglês para usar a Internet. Para diminuir essa barreira, alguns provedores de informação e servidores WWW mantêm múltiplas cópias de documentos em línguas diferentes. Entretanto, devido à natureza dinâmica da WWW, a manutenção de cópias multilinguais é problemática, podendo acarretar inconsistência de dados e dificuldades no gerenciamento dos documentos. Uma outra alternativa que já está sendo bastante difundida é a da utilização de tradutores automáticos, que podem se basear em três técnicas distintas, a saber: a direta, a de transferência e a interlíngua. Na tradução direta, rotinas de substituição e reconhecimento/casamento de padrões altamente específicas são utilizadas para permitir a troca de palavras ou estruturas da língua fonte por palavras ou estruturas correspondentes da língua destino. Nos sistemas de transferência obtém-se uma estrutura representativa da sentença na língua original (por meio de um processo de análise) e é esta estrutura que é transferida aos padrões da língua destino, para então ser sintetizada na forma textual. Já na abordagem interlíngua, não há a necessidade de módulos particulares para cada par de línguas envolvido na tradução. Somente dois estágios são necessários: o de mapeamento de sentenças da língua fonte em uma representação neutra em relação às línguas envolvidas no processo e o de linearização dessa representação neutra em sentenças da língua destino, o que torna a abordagem de interlíngua a alternativa mais atraente para sistemas multilinguais. Esse conceito já vem sendo adotado em diversos projetos de pesquisa em tradução automática (p.ex., [1], [2]), e tem a vantagem de requerer um número menor de sistemas. Por exemplo, para n línguas, são necessários $2n$ sistemas (n para codificação para a interlíngua e n para decodificação da interlíngua para as línguas naturais), ao passo que as outras abordagens requerem $n(n-1)$ sistemas. Sua maior vantagem, entretanto, está na uniformidade alcançada com uma representação do conhecimento que será comum para todas as línguas.

A abordagem da interlíngua foi a escolhida para o projeto *Universal Networking Language* (UNL), uma iniciativa do IAS/UNU (*Institute of Advanced Studies /United Nations*

University), sediado em Tóquio. Preocupada com a barreira da língua para comunicação na Internet em nível mundial, a UNU está patrocinando um projeto que visa ao desenvolvimento de codificadores e decodificadores para uma interlíngua também denominada *Universal Networking Language* (UNL), desenvolvida por pesquisadores japoneses da UNU [3,4]. O projeto tem duração prevista de dez anos, a partir de 1997, e hoje inclui 15 línguas: árabe, alemão, chinês, espanhol, francês, hindi, indonésio, inglês, italiano, japonês, letão, mongol, português, russo e tailandês. O objetivo final é que cada usuário da Internet possa produzir seu material na sua própria língua, disponibilizá-lo em UNL, através do codificador para UNL. Um outro usuário poderá então receber a informação numa outra língua, empregando o decodificador de UNL para a sua língua. 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. Os sistemas para o português estão sendo desenvolvidos sob a coordenação geral do Sr. Tadao Takahashi (Projeto UNL-Brasil). As ferramentas de PLN, propriamente ditas, estão a cargo do Núcleo Interinstitucional de Lingüística Computacional (NILC), que conta com docentes e pesquisadores da USP de São Carlos, da Universidade Federal de São Carlos (UFSCar) e da Unesp de Araraquara[1].

O primeiro sistema que está sendo desenvolvido é o decodificador para as várias línguas. Este artigo apresenta a metodologia adotada na implementação do decodificador para o português, incluindo os módulos necessários para seu funcionamento. Para tanto, introduzimos resumidamente a proposta da linguagem UNL na próxima seção. O processo e a ferramenta adotada para a decodificação de UNL para português, bem como a metodologia adotada, são descritos na Seção 3. Resultados de decodificação são apresentados na Seção 4, sendo que a Seção 5 traz as conclusões e aponta para os trabalhos futuros no Projeto UNL.

2. A linguagem UNL

A UNL foi concebida para representar de forma única o conteúdo semântico de uma sentença escrita em qualquer língua natural. Mais especificamente, a UNL é uma metalinguagem que serve para descrever aspectos especiais do significado de sentenças, tais como as relações semânticas que podem ser representadas por relações formais (morfológicas ou sintáticas) entre palavras de uma sentença [5]. A UNL se assemelha a outras metalinguagens definidas no âmbito da lingüística computacional, da psicologia ou da lingüística, com respeito à descrição do relacionamento semântico entre conceitos. Por exemplo, Schank formalizou as relações semânticas representativas do significado em sua Teoria de Dependência Conceitual [6]; Kintsch estabeleceu um modelo semântico-representacional mental [7], enquanto Fillmore [8] e Jackendoff [9] tentaram descrever as relações semânticas sistematicamente. Apesar de simples, a UNL fornece uma visão muito prática de um sistema de PLN: a de permitir o tratamento de aspectos semânticos de forma sistemática e independente do conhecimento profundo sobre questões não textuais. Ela se restringe, por exemplo, ao significado literal (e, logo, denotativo) de sentenças e trata de descrições de significados frasais seguindo a abordagem tradicional de relacionamento de papéis semânticos a entidades ou objetos componentes da frase. Ela não abrange alguns pressupostos teóricos importantes para a comunicação, tais como as questões comunicativas, estilísticas, intencionais ou retóricas, que levam ao significado conotativo, de natureza pragmática. Desse modo, suas limitações indicam também a exclusão do tratamento da recepção da linguagem, tida no Projeto UNL como um problema exclusivo do leitor.

[1] http://www.nilc.icmc.sc.usp.br

Apesar dessas limitações, a UNL permite reconhecer algumas expressões idiomáticas, que são lexicalizadas com base em interpretações literais. Este é o elo mais explícito, na UNL, da tentativa de se processar informações de caráter conotativo. Entretanto, a UNL visa a promover somente a comunicação básica entre indivíduos de línguas nativas distintas e não a comunicação natural e abrangente que existiria entre indivíduos que partilhassem a mesma língua. A característica principal da UNL é a de privilegiar a predicação entre elementos lingüísticos, por sua caracterização individual ou relacional (com outros elementos lingüísticos) durante a ocorrência de eventos ou outras relações complexas entre eventos. Neste caso, os próprios argumentos verbais ou adjuntos frasais se manifestam como predicados, sendo responsáveis por preservar a relevância da informação textual. As predicações podem, portanto, ser relativamente simples, como um simples adjetivo, ou complexas, como em uma subordinação.

Privilegiando os aspectos estruturais do significado, em vez dos aspectos da semântica lexical, a UNL se baseia em conceitos relacionados a sistemas de *relações primitivas semanticamente universais* [5]. A semântica lexical é incorporada ao léxico, como um ponto de partida para o processamento subseqüente. Desse modo, parte da representação semântica já é recuperada pelas próprias entradas lexicais. Em especial, incorpora-se ao léxico as interpretações particulares da língua em cada um de seus verbetes e são esses que passam a corresponder ao vocabulário UNL. A UNL serve, portanto, para projetar entidades de sistemas de representação da estrutura gramatical em sistemas de representação de interpretações semânticas e vice-versa. Por exemplo, no processo de codificação, ela é usada para mapear uma sentença escrita em língua natural (sentença de origem) em um conjunto de relações entre significados (sentença UNL), sendo que cada uma das relações consiste em uma proposição no sistema de representação universal. No processo de decodificação, regras gramaticais de uma língua destino qualquer são aplicadas à sentença UNL, para gerar a estrutura gramatical correspondente, a partir da qual se obtém a forma superficial na língua destino.

São três os principais constituintes da UNL: i) Palavras Universais (*Universal Words* - UWs), i.e., palavras que retêm o significado dos componentes frasais; ii) Rótulos de relações (*Relation Labels* - RLs), que expressam as relações entre as UWs; iii) Rótulos de atributos (*Attribute Labels* - ALs), que expressam informações adicionais e, em geral, restritivas, sobre as UWs.

A função de uma UW é denotar um significado específico. Sua representação genérica é um rótulo simples (que indica o significado genérico de uma palavra em inglês) ou um rótulo limitado por um intervalo específico, que denota significados distintos (quando há, p.ex., ambigüidade em relação à palavra original). Por exemplo, "book" permite a representação das seguintes UWS: *book*, *book(icl>publication)*, *book(=account)* e *book(obj>room)*. A primeira UW, *book*, é a representação mais genérica do significado. As demais limitam este significado a conceitos particulares. No caso, a *publicações*, *livro de contabilidade* ou *reserva de um quarto* (em um hotel), respectivamente.

RLs servem para expressar relações binárias entre significados, i.e., entre duas UWs distintas. Sua representação geral é dada por um par ordenado do tipo *relation_label(UW1, UW2)*, onde UW_1 e UW_2 são duas UWs diferentes relacionadas pela relação semântica indicada por *relation_label*. Há diversas classes de RLs, como algumas apresentadas a seguir, perfazendo um total de 40 RLs.

RLs entre componentes de uma sentença:

Agent (agt): "um agente que causa uma ação volitiva" ("agente" é um objeto animado com intenções). Exemplo: *A lebre corre*, representada em UNL por *agt(run.@present,hare.@def)*.

Object (obj): "um objeto de uma ação ou mudança que afeta o objeto. Exemplo, *Pedro come maçãs,* representada por
agt(eat.@present,Pedro),
obj(eat.@present,apple.@generic.@pl).

Outros rótulos de relação entre componentes sentenciais incluem objeto atributivo, método, tempo, possuidor, beneficiário.

RLs entre UWs:

Inclusion (icl), para representar, p.ex., hiperonímia, como em *icl(cão, animal)*, ou meronímia (relação parte-todo), como em *icl(braço,corpo_humano).*

Synonym (= ou *equ*), para representar significados equivalentes entre UWs, como em
=(livro, livro_comercial).

ALs servem para limitar o significado de uma UW genérica, i.e., para particularizar seu significado. Informações adicionais tais como tempo verbal, aspecto, intenção ou estrutura sentencial são exemplos de atributos específicos de uma UW. A representação genérica de um AL é dada pela UW, seguida por tantos atributos quantos forem os sugeridos na sentença de origem. Cada um deles é identificado pelo símbolo inicial "@". Por exemplo, a representação genérica de uma UW com *n* atributos tem a forma:

$$UW.@atrib_1.@atrib_2.....@atrib_n$$

Se não houver ALs vinculados a uma UW, esta representa o significado mais genérico de sua classe. Da mesma forma que para RLs, pode haver diferentes classes de ALs. Apresentamos, a seguir, algumas delas.

• **ALs que limitam o poder de expressão de uma UW:**

@generic: indica uma UW geral. Exemplo *Pedro come maçãs*, com as ALs restritivas: *@pl* e @generic, representadas em UNL por:
agt(eat.@present,Pedro),
obj(eat.@present,apple.@generic.@pl).

@def: indica uma UW que já foi referenciada anteriormente. Exemplo:
Sentença original: *Pedro fechou a porta.*
Representação em UNL: *agt(close.@past,Pedro),*
obj(shut.@past,door.@def)

Outros rótulos de atributos que limitam o poder de expressão de UWs incluem *@indef,* para UWs indefinidas, "não específicas, mas não genéricas", *@pl,* "plural, mas não genérica", e *@not,* representando informações complementares, no sentido lógico.

• **ALs que expressam tempo verbal.** Estes seguem a gramática da língua inglesa e incluem:
@past: "evento no passado".
@present: "evento no presente".
@future: "evento no futuro".

• **ALs que expressam aspecto:**

@begin-soon: "evento que vai começar". Exemplo:
Sentença original: *O avião está para aterrissar.*
Representação em UNL: *agt(land_in.@begin-soon,plane.@def).*

@begin-just: "evento recém-começado". Exemplo:
Sentença original: *O jogo acabou de começar.*
Representação em UNL: *obj(start.@begin-just,game.@def).*

Outros exemplos de rótulos de atributos que expressam aspecto incluem *@end-soon,* para "eventos que estão quase terminando", *@end-just,* "eventos recém-terminados", *@progress,* eventos progressivos e *@repeat,* repetição de um mesmo evento, envolvendo o mesmo agente/objeto.

• **ALs que expressam intenções:**

@focus: Como em *Foi você quem saiu?*,

Representação em UNL: *agt(leave.@entry.@interrogation.@past, you.@focus)*

Outros rótulos de atributo de UWs que expressam intenções podem ser: *@emphasis, @topic, @intention, @recommendation*, etc.

Alguns dos ALs descritos acima se aplicam a elementos intra-sentenciais, outros à sentença como um todo. No primeiro caso, o AL é associado ao componente específico que ele modifica. No segundo, o AL é associado à UW nuclear, i.e., àquela que expressa o predicado principal da sentença. Formas mais complexas de combinações de UWs podem ainda ser expressas por meio de ALs. Além disso, línguas naturais distintas podem ter representações mais detalhadas para expressar informações restritivas sobre conceitos semânticos. Neste caso, a UNL permite a definição de subcategorizações pela inclusão de novos atributos e, assim, é possível tratar as particularidades de cada língua.

2.4. Exemplo de sentença UNL

Uma sentença UNL consiste no inter-relacionamento entre UWs oriundas do significado expresso na sentença original, expresso por meio de RLs. Este significado, ou o significado de componentes da sentença UNL, pode ser restringido com o auxílio dos ALs. Uma sentença UNL é ilustrada na Figura 1, que corresponde à sentença em português:

Há muito tempo, na cidade de Babilônia, o povo começou a construir uma torre imensa, que parecia alcançar os céus.

```
[S]
  tim(begin(icl>event).@entry.@pred.@past, long_ago)
  nam(city(icl>place).@def, Babylon(icl>city))
  ppl(begin(icl>event).@entry.@pred.@past, city(icl>place).@def)
  agt(begin(icl>event).@entry.@pred.@past, people(icl>human).@def)
  obj(begin(icl>event).@entry.@pred.@past,
     build(equ>construct,agt>human,obj>structure).@pred)
  agt(build(equ>construct,agt>human,obj>structure).@pred,
     people(icl>human).@def)
  obj(build(equ>construct,agt>human,obj>structure).@pred,
     tower(icl>building).@indef)
  aoj(huge(aoj>entity), tower(icl>building).@indef)
  aoj(seem(icl>event).@pred.@past, tower(icl>building).@indef)
  obj(seem(icl>event).@pred.@past,
     reach(gol>entity,obj>entity).@pred.@begin-soon)
  obj(reach(gol>entity,obj>entity).@pred.@begin-soon, tower(icl>building).@indef)
  gol(reach(gol>entity,obj>entity).@pred.@begin-soon, heaven(ant>hell).@def.@pl)
[/S]
```

Figura 1: Exemplo de Sentença UNL

3. O decodificador UNL-português

No Projeto UNL, o IAS/UNU forneceu um interpretador de regras - ou decodificador - chamado DeCo [10]. O NILC adotou o DeCo como núcleo do sistema de processamento do português para decodificar sentenças UNL em sentenças em português. Para a geração a partir do DeCo, informações léxico-gramaticais específicas da língua portuguesa devem ser incluídas. O DeCo é uma ferramenta que implementa uma máquina de estados cujo objetivo é decodificar uma dada sentença UNL em uma sentença escrita em uma língua destino (Figura 2). O DeCo compreende basicamente dois processos: a resolução dos relacionamentos

semânticos entre UWs que compõem a sentença UNL, vistos como uma rede de nós (*nodenet*), juntamente com a resolução de seus atributos gramaticais, e o controle sobre janelas de uma lista de nós (*nodelist*), que contém informações a serem gradualmente modificadas por operações expressas em regras, até que a sentença na língua destino seja completamente gerada. O DeCo utiliza um dicionário UNL-língua destino com as associações entre palavras e expressões da língua natural e o conteúdo semântico veiculado pelas UWs . Além disso, baseia-se num conjunto de regras de mapeamento UNL-língua destino, que produzirão as modificações necessárias na *nodelist*, a fim de gerar uma sentença naquela língua. O decodificador é, portanto, um sistema de geração de linguagem, que é usado para produzir uma sentença numa determinada língua. A seguir são apresentadas as características dos três módulos da Figura 2.

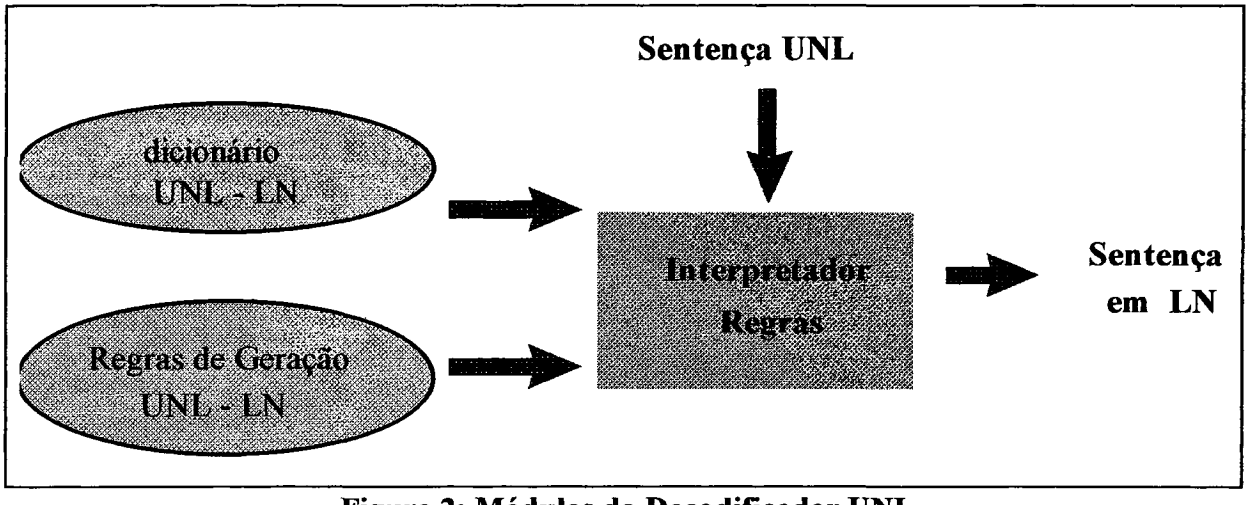

Figura 2: Módulos do Decodificador UNL

3.1. O Dicionário UNL-português

O dicionário da UNL para o português do Brasil inclui headwords para cerca de 63000 UWs, representando uma cobertura razoável do vocabulário básico da língua [11]. A sintaxe do dicionário UNL-português obedece o seguinte formato:

[headword] canônica "UW" (grammatical features) <P,f,p>;

onde: *headword* é a palavra do português, correspondente ao significado expresso pela UW;

"canônica" é a forma canônica da palavra em português;

(*grammatical features*) é o conjunto de traços gramaticais e semânticos da *headword* (informação de natureza morfológica, por exemplo, gênero e número);

P: denota português;

f e **p** são valores naturais que exprimem a freqüência e a prioridade de uso da *headword,* utilizados, respectivamente, para codificação português-UNL e decodificação UNL-português.

As entradas da Figura 3 ilustram o conteúdo de tal dicionário antes do processo de associação com o português. Todas as entradas derivadas de uma única palavra em inglês correspondem às suas diferentes acepções. Vale notar as restrições semânticas impostas à UW mais genérica "*threaten*" em cada caso, originando diferentes acepções. Estas restrições envolvem RLs (p.ex., *agt*, *obj*) e outras UWs (p.ex., *danger*, *human*, *trouble*).

```
[]{} "threaten" ();
[]{} "threaten(agt>human,obj>danger)" ();
[]{} "threaten(agt>human,obj>entity)" ();
[]{} "threaten(agt>human,obj>human)" ();
[]{} "threaten(agt>human,obj>trouble)" ();
[]{} "threaten(icl>event)" ();
[]{} "threaten(icl>event,obj>human)" ();
```

Figura 3: Exemplos de Entradas do Dicionário UNL

Em vista das características do DeCo, adotamos os radicais morfológicos como forma de representação das *headwords* em português, para restringir o número de entradas dicionarizadas. O principal impacto dessa decisão está na representação da classe de verbos, cujo processo flexional passa a ocorrer via regras. Entretanto, em várias situações, não se pôde manter a consistência morfológica da análise, seja (1) pela necessidade de regularizar todas as ocorrências verbais (o que fez com que, em muitos casos, dicionarizássemos apenas a parte dos radicais que precederia a formação dos alomorfes (caso de [fi], representando o verbo "ficar", que poderia se realizar como [fic] ou [fiqu]; e (2) pela necessidade de representar as formas irregulares de substantivos e adjetivos no plural (caso de [ação] e [ações], ambas dicionarizadas).

3.2. O Interpretador de Regras

Uma vez escolhida uma UW para ser processada, o dicionário é pesquisado antes da aplicação de qualquer regra de geração, a fim de associar à UW informações relativas à *headword* correspondente e a seus atributos gramaticais. Somente após a obtenção dessas informações o DeCo inicia a busca por uma regra de geração que contemple as informações da *nodelist* que estão visíveis em suas janelas. As regras de geração são aplicadas apenas aos nós da *nodelist*, a partir de sua configuração atual e da configuração da *nodenet*. O processo de geração ocorre sob controle de dois tipos de janelas: de geração (G) e de condição (C) (Figura 4). As janelas de geração à esquerda e à direita destinam-se a examinar os atributos gramaticais de seus respectivos nós e, portanto, decidir qual a alteração de conteúdo que a *nodelist* pode sofrer. As janelas de condição servem para examinar as condições dos nós da *nodelist* que são vizinhos das janelas de geração e, portanto, servem para verificar as condições de contexto no processo de geração. Os nós das janelas de condição são examinados também com base em suas características gramaticais.

Ao ter acesso aos nós vizinhos de ambos os lados das janelas de geração, as janelas de condição permitem modificações adicionais da *nodelist*, fornecendo informações sobre a possibilidade de aplicação de uma dada regra de geração. Juntos, esses dois tipos de janela provêm informações para a aplicação de uma regra específica de geração, em função do estado atual da *nodelist*. Quando aplicada, a regra produz uma das seguintes alterações na *nodelist* (não exclusivas): a) atualização de conteúdo dos nós envolvidos (inserção ou eliminação de alguns de seus atributos gramaticais); b) inserção de um novo nó (oriundo da *nodenet* ou não) à esquerda ou à direita das janelas de geração. Após uma dessas alterações, um movimento à esquerda ou à direita das janelas de geração também pode ocorrer.

A Figura 4 mostra o status da *nodelist* e da *nodenet* no início do processo de decodificação da sentença UNL que envolve os seguintes RLs (a notação a seguir é uma forma simplificada que suprime detalhes desnecessários nesta ilustração):

agt([UW=build], [UW=people]);
obj([UW=build], [UW=tower]);
aoj([UW=tower], [UW=huge]);

Após decodificada para o português, esta sentença dá origem a *"O povo construiu uma torre imensa"*. A inserção de nós na *nodelist* que não estão aparentes na *nodenet* tem o objetivo de descrever condições especiais de uma UW, como, por exemplo, sufixos que sinalizam tempo verbal a partir de radicais (p.ex., inserção de "uiu" no radical "constr"). Este tipo de inserção é sinalizado por um trigger gramatical representado por "!". Por exemplo, *!present* indica a geração do item lexical derivado de um verbo na forma infinitiva, pela adição de características relativas ao tempo "presente". Em geral, podemos sumarizar o processo de aplicação de regras de geração no DeCo da seguinte forma: a inclusão de nós na *nodelist* é feita com base nos seguintes parâmetros:

a) no conteúdo das janelas de geração à direita e à esquerda;
b) nos nós vizinhos destas janelas;
c) no nó da *nodenet* que é candidato a ir para a *nodelist*. Neste caso, o nó [UW=build] do exemplo anterior seria o primeiro a ser transferido da *nodenet* para a *nodelist*. Essa prioridade é indicada pelo AL *@entry*. Nota-se, na Figura 4, a complementação (também simplificada) das palavras ou radicais em português correspondentes às UWs *povo*, *torre* e *imens*. Essa complementação ocorre previamente à inserção na *nodelist* (o nó da *nodelist* da Figura 4 já teria sofrido alterações, passando da *headword "constr"* para a palavra *"construiu"*, após a resolução dos *Attribute Labels*).

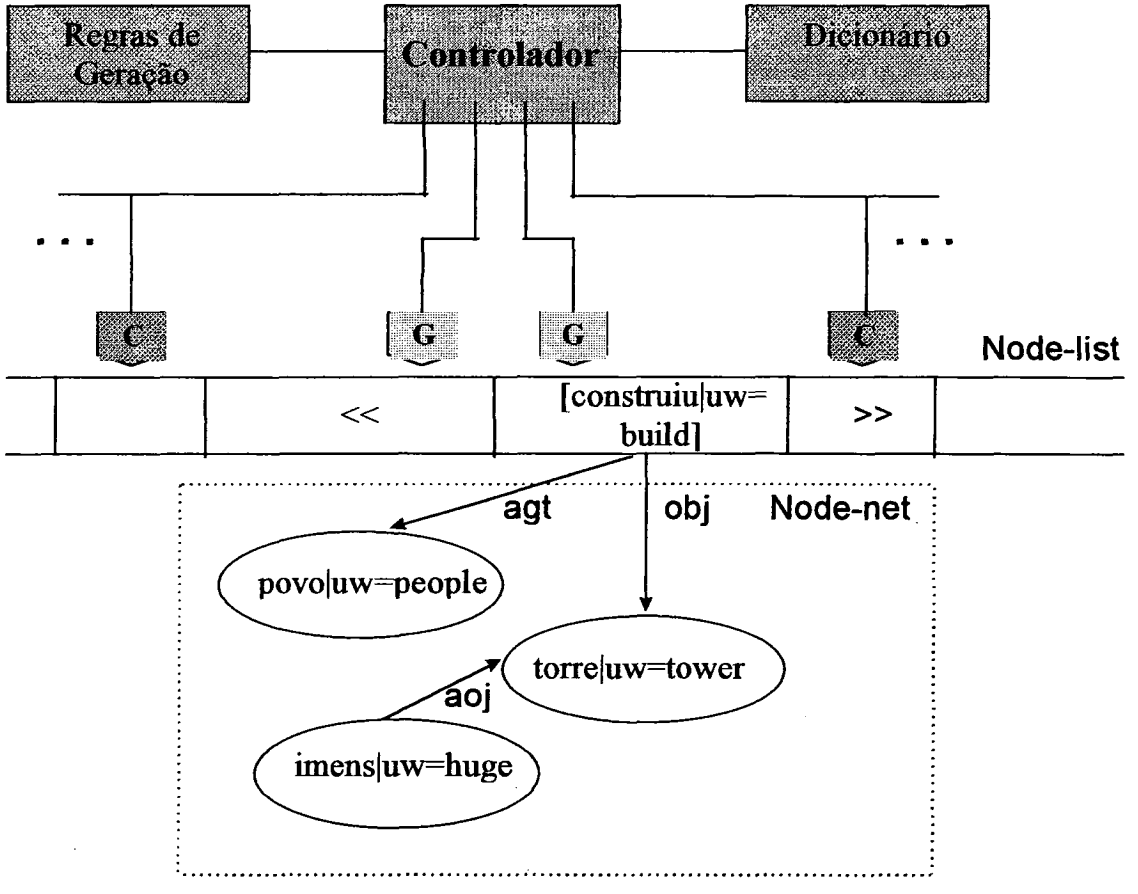

Figura 4: Configuração do DeCo para "O povo construiu uma torre imensa"

3.3. Regras de Geração UNL-português

De acordo com o que foi descrito, a sintaxe das regras de geração que alimentam o DeCo compreende três diferentes padrões, especificados a seguir (M pode ser L, R, : ou ?, correspondendo, respectivamente, a movimento à esquerda, movimento à direita de ambas as janelas de geração, alteração de atributos gramaticais dos nós sob as janelas de geração e operação de *backtracking*).

1) Modificação de Atributos.

{condição:ação:relação:função}M{condição:ação:relação:função}

2) Inserção (de n_j) à esquerda de qualquer janela de geração.

"condição:ação:relação:função"M{condição:ação:relação:função}

onde *relação* associa o nó n_i sob uma das janelas de geração a um nó n_j na *nodenet* e n_i e n_j satisfazem as condições de aplicabilidade da regra de geração.

3) Inserção à direita de qualquer janela de geração.

{condição:ação:relação:função}M"condição:ação:relação:função"

Esta operação é análoga à anterior, mas o nó da *nodenet* é inserido à direita da janela.

Um processo de *backtracking* pode ocorrer tanto sobre a escolha de uma *headword* como sobre as regras de geração (retornando a um estado prévio para a busca de uma regra alternativa), mas nunca desfaz uma operação de inserção de um nó na *nodelist*. A associação de informações obtida através das janelas de condição e geração permite o tratamento de dependência de contexto.

A formulação de regras de decodificação UNL-português para a utilização da ferramenta de decodificação DeCo foi baseada num estudo exaustivo das manifestações morfossintáticas associadas às RLs e ALs para o português do Brasil [12]. As Tabelas 1 e 2 mostram alguns exemplos de mapeamento entre RLs e Als no português, respectivamente. Em [13, 14] o leitor encontra mais detalhes sobre o conjunto de 5500 regras (a maioria delas trata de aspectos morfológicos da geração).

Tabela 1: Exemplos de Mapeamento entre RLs e Manifestações Morfossintáticas

RLs	Categorias Sintáticas mais freqüentes	Categorias morfossintáticas
Soj	sujeito	verbos de ligação (ou não) - substantivos concretos ou abstratos
Obj	objeto direto	verbo - substantivo comum em posição de objeto.
agt	sujeito	verbo - substantivo animado (substituído ou não por um pronome pessoal
aoj	adjunto adnominal	substantivo - adjetivo
tim	adjunto adverbial de tempo	verbo - advérbio ou locução adverbial.
mod	adjunto adnominal e adjunto adverbial	entre diversas classes de palavras
ppl	adjunto adverbial	verbo - adjunto adverbial
opl	objeto direto	verbo - substantivo
pos	complemento nominal	substantivo - sintagma preposicional
seq	coordenação	dois verbos em diferentes sentenças

gol	variável: objeto direto ou indireto, complemento nominal, coordenação	Não há linearidade sintática
man	adjunto adverbial	verbo - advérbio ou locução adverbial
ptn	objeto indireto	verbo - locução promominal
lpl	adjunto adverbial	sintagma preposicional - substantivo
qua	adjunto adnominal	numeral - substantivo
bas	adjunto adverbial	advérbio - substantivo ou verbo

Tabela 2: Exemplos de Mapeamento entre ALs e Manifestações Lingüísticas

ALs	Função	Manifestação Lingüística
entry	marca nó principal de orações simples ou hierarquia entre orações de uma sentença	núcleo do predicador: verbo, núcleo do predicado verbal ou predicativo do sujeito em orações com verbo ser. Em períodos compostos, o verbo da oração que exprime conseqüência em relação à proposição que lhe antecede.
present, past, future	indicam as desinências temporais	predicados verbais ou estruturas com predicado nominal (quando recaem sobre nomes)
pred	sinaliza uma UW predicativa	núcleo de um sintagma verbal ou núcleo de um predicado nominal
begin-soon	sinaliza um evento que está prestes a acontecer	advérbio de tempo
apodosis	incide sobre a conclusão de uma condição	frase condicional
ability	expressa noção modal de habilidade e capacidade	verbo "poder" e "ser capaz de"
state	indica evento concluído com resultado permanente	pretérito perfeito do indicativo
progress	indica evento em progresso	perífrase verbal estar + verbo principal na forma gerundiva (com sufixo -ndo)
complete	indica evento concluído	verbo no pretérito perfeito do indicativo
def	sinaliza um artigo definido	artigos definidos
indef	sinaliza um artigo indefinido	artigos indefinidos
pl	indica desinência de número	morfema -s e seus alomorfes
sub	marca dependência temporal	elementos coordenados de um sintagma nominal
not	indica negação	negação de um verbo ou predicadores de negação de um item lexical

4. Resultados de geração

O sistema de geração para o português do Brasil desenvolvido no âmbito do projeto UNL permite a linearização de texto com as mais variadas estruturas sintáticas. Em condições ideais, os resultados são bastante promissores, pois a mensagem é transmitida com precisão, como ilustra o exemplo abaixo, em que a sentença original em inglês foi codificada manualmente para UNL e então decodificada com o DeCo.

Sentença original em inglês:
It shall function in accordance with the annexed Statute, which is based upon the Statute of the Permanent Court of International Justice and forms an integral part of the present Charter.

Saída do Deco em Português:
A corte funcionará de acordo com o estatuto anexo que se baseia no estatuto da corte permanente de justiça internacional e constitui uma parte integrante da carta presente.

Por condições ideais, queremos dizer que a codificação foi precisa, com as *headwords* em português sendo escolhidas a partir de UWs – já tratadas com as restrições semânticas - selecionadas também com precisão. Acrescente-se que, a despeito da sofisticação sintática da sentença, ela não apresenta dificuldades no que concerne à dependência de informações contextuais, e nem ambigüidades. O desempenho geral do sistema quando utilizado na Internet para propósitos gerais, como é o objetivo do projeto, deverá ser bastante inferior. Em primeiro lugar, estudos recentes [15] indicaram que a codificação – mesmo realizada por especialistas humanos – pode ter uma dependência da língua natural, da sentença original. Ou seja, alguma informação pode ser perdida já na codificação. Além disso, a escolha das UWs com as restrições corretas não é trivial, e uma correta seleção de headwords depende da disponibilidade de um dicionário enorme – existem cerca de 1 milhão de UWs já catalogadas. Note, por exemplo, como a qualidade da saída para a sentença abaixo é prejudicada pela inadequação do dicionário de UWs:

Sentença original em inglês:
In the final game, the spectators had to wait until the 70th minute for the first goal to be scored: Antonin Puc sent the Czech team into the lead.

Saída do Deco em Português:
Os espectadores tiveram que esperar até o minuto de 70th no jogo final para ser marcado o primeiro gol: Antonin Puc enviou o time tcheco para a liderança.

Neste exemplo, foi utilizada uma versão do dicionário que não continha a palavra septuagésimo e em que a UW correspondente ao verbo "send" com significado de "colocar" não havia sido tratada. A limitação maior do sistema de geração, todavia, está na impossibilidade de tratar fenômenos supra-sentenciais, pois a UNL só trata o texto como uma soma de sentenças isoladas.

5. Comentários Finais

O Projeto UNL constitui-se num esforço da Universidade das Nações Unidas para, no longo prazo, minimizar a barreira da língua na comunicação internacional via Internet. Com respeito à metodologia de desenvolvimento/emprego da UNL como interlíngua, o Projeto UNL possui uma abordagem tradicional para descrever expressões lingüísticas, que é a combinação de palavras de um vocabulário universal pelo uso de regras que permitem a codificação e decodificação de expressões UNL bem formadas. Entretanto, quando comparada às linguagens formais clássicas, que usam algum critério de ordem de análise

(p.ex., *left-to-right*) ou relações formais entre unidades lingüísticas, a UNL possui a vantagem de constituir uma descrição prática de alguns aspectos cruciais do significado de sentenças, fazendo a correspondência entre relações semânticas e relações formais em nível de morfologia ou sintaxe e, logo, em nível de processamento da estrutura superficial das sentenças [5]. Um outro aspecto da viabilidade do uso da UNL diz respeito à ênfase ao processamento do significado literal expresso textualmente, que a isola dos problemas cruciais do tratamento de aspectos pragmáticos da comunicação, embora restrinja seu poder comunicativo. Lembramos, no entanto, que o principal objetivo da UNL é justamente a comunicação *básica*, possível no contexto de representação delimitado pela UNL. Ressaltamos, ainda, que aplicativos extremamente úteis podem ser desenvolvidos no curto prazo, empregando a tecnologia UNL. Em especial, aplicativos cujos problemas de compreensão podem ficar a cargo do leitor. Por exemplo, ferramentas para codificação e decodificação de *homepages*, cujo conteúdo mais genérico normalmente independe de contextualização, além de sistemas de sumarização automática e indexação de grandes quantidades de texto em diversas línguas.

De outro lado, os sistemas de codificação e decodificação são, obviamente, extremamente dependentes da disponibilidade de um léxico adequado. A incorporação da semântica lexical ao dicionário de UWs-português em nosso projeto substitui, de forma bastante rudimentar, a necessidade de se interpretar a linguagem de forma conotativa. Neste caso, podemos considerar que o léxico já incorpora parte da associação de regras gramaticais ao vocabulário utilizado na interação. Do ponto de vista do caráter internacional do Projeto UNL, esta perspectiva é bastante interessante, particularmente no contexto de comunicação via Internet, pois ela sugere o encapsulamento das questões lingüísticas problemáticas ao processamento de cada língua natural particular.

Agradecimentos

O desenvolvimento do sistema de decodificação foi parcialmente patrocinado pela Universidade das Nações Unidas. Os autores também agradecem a toda a equipe do NILC e ao CNPq pelo apoio na forma de bolsas de pesquisadores.

Referências

[1] Bian, G.; Chen, H. (1997). *An MT Meta-Server for Information Retrieval on WWW.* Proc. of the Natural Language Processing for the World Wide Web. 1997 AAAI Spring Symposium Series. Stanford University. USA.

[2] Zajac, R.; Casper, M. (1977). *The Temple Web Translator.* Proc. of the Natural Language Processing for the World Wide Web. 1997 AAAI Spring Symposium Series. Stanford University. USA.

[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.

[5] Dillinger, M. (1997). *The Universal Networking Language Project: Principles, Perspectives and Current Work.* Projeto UNL/Brazil: II Encontro de Trabalho. NCE/UFRJ, Rio de Janeiro - RJ, 13-14 de Agosto.

[6] Schank, R. (1975). *Conceptual Information Processing*. North-Holland Publishing Company.

[7] Kintsch, W. (1974). *The Representation of Meaning in Memory*. Erlbaum. Hillsdale, NJ.

[8] Fillmore, C. (1968). The Case for Case. In E. Bach and R. Harms (eds.), *Universals in Linguistic Theory*. Holt, Rinehart & Winston. New York.

[9] Jackendoff, R. (1983). *Semantics and Cognition*. MIT Press. Cambridge, MA.

[10] UNL (1997). *DeConverter Specification*. Version 1.2 (Tech. Rep. UNL-TR1997-010). UNU/IAS/UNL Center. Tokyo, Japan.

[11] Dias da Silva, B.C.; Sossolote, C; Zavaglia, C.; Montilha, G.; Rino, L.H.M.; Nunes, M.G.V.; Oliveira Jr., O.N.; Aluísio, S.M. (1998) The design of the Brazilian Portuguese machine tractable dictionary for an interlingua sentence generator. III Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR' 98). Porto Alegre - RS. Novembro, p.71-78.

[12] Sossolote, C.R.C., Zavaglia, C., Rino, L.H.M. and Nunes, M.G.V.(1997). *As Manifestações Morfossintáticas da UNL no português do Brasil*. Notas do ICMC-USP, 36, Novembro 1997 (Tech. Rep. NILC-TR-97-2).

[13] Martins, R.T.; Rino, L.H.M., Nunes, M.G.V. (1998a). *As Regras Gramaticais para a Decodificação UNL-Português no Projeto UNL*. Relatório Técnico 67. Instituto de Ciências Matemáticas e da Computação. Universidade de São Paulo, São Carlos.

[14] Martins, R.T.; Rino, L.H.M.; Nunes, M.G.V.; Oliveira Jr., O.N. (1998b). Can the syntactic realization be detached from the syntactic analysis during generation of natural language sentences? III Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR' 98). Porto Alegre - RS. Novembro, p.32-37.

[15] Martins, R.T.; Rino, L.H.M.; Nunes, M.G.V.; Montilha, G.; Oliveira Jr., O.N. (2000). An interlingua aiming at communication on the Web: How language-independent can it be? Proceedings of the Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP, pp. 24-33. NAACL-ANLP 2000 Workshop, April. Seattle, Washington, USA.

NOTAS DO ICMC

SÉRIE COMPUTAÇÃO

060/2001 SILVA, E.Q.; MOREIRA, D.A. – Use of software agents to the management of distance education courses over the internet.

059/2001 OLIVEIRA, M.C.F.; LEVKOWITZ, H. – Visual data exploration and mining: a survey.

058/2001 SOARES, M D.; FORTES, R P M; MOREIRA, D A – Version–web : a tool for helping web pages version control.

057/2001 LIANG, Z; MACAU, E E N; OMAR, N - Scene Segmentation of the Chaotic Oscillator Network.

056/2000 BATISTA, G E A P A; CARVALHO, A C P L F; MONARD, M C – Applying one-sided selection to unbalanced datasets.

055/2000 NONATO, L G; MINGHIM, R.; OLIVEIRA, M C F; TAVARES, G. – A novel approach for delaunay 3D reconstruction with a comparative analysis in the light of applications.

054/2000 MORSELLI JR., J C M; SANTANA, R H C; SANTANA, M J; ULSON, R S – An approach for dynamic swapping of distributed simulation synchronisation protocols.

053/2000 SPOLON, R.; SANTANA, M J; SANTANA, R H C – A methodology for performance evaluation of optimistic distributed simulation synchronisation mechanisms.

052/2000 BRANCO, K R L J C; SANTANA, M J; SANTANA, R H C; CALÔNEGO JUNIOR, N – A parallel programming supporting tool.

051/2000 FRANCÊS, C R L; VIJAYKUMAR, N L; SANTANA, M J; CARVALHO, S V de; SANTANA, R H C – Stochastic statecharts for obtaining performance measurements of a file server model.