

The decoding system for Brazilian Portuguese using the Universal Networking Language(UNL)

Por João Luiz Martelli Moreira - Fpolis, Maio de 2002

Créditos:

Maria das Graças V. Nunes

Ronaldo T. Martins

Lúcia H.M.Rino

Oswaldo N. Oliveira Jr.

Núcleo Interinstitucional de Lingüística

Computacional - NILC

ICMC-USP - UFSCar, São Carlos, SP, Brazil

Abstract

- Uso do DoCo da UNL-Center, com regras e dicionário para Português-Brasil;
- Procedimentos para desenvolver o dicionário e regras de geração;
- Resultados promissores para sentenças complexas, permitindo captar a essência dos textos;
- Limitações do DeCo

Introdução

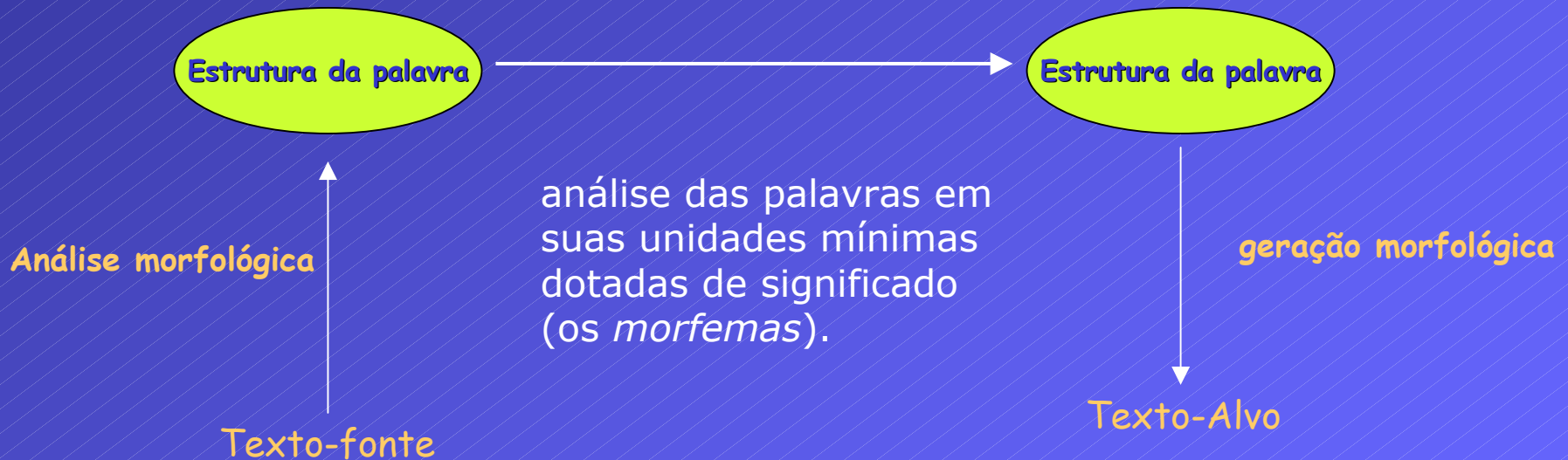
- A UNL foi concebida para minimizar as barreiras na comunicação global;
- Trata-se de uma aplicação que permite a conversão de uma linguagem natural para um conjunto de relações semânticas, a qual, posteriormente, permite uma nova conversão para linguagem natural em outro idioma;
- Apresentar metodologia para implementar um "*decoding*" para português-Brasil.

A Interlíngua UNL

- Formalismo capaz de representar um subconjunto semântico de sentenças escritas em linguagem natural;
- Meta-linguagem capaz de representar o significado dos aspectos literais de uma sentença.

A Interlíngua UNL

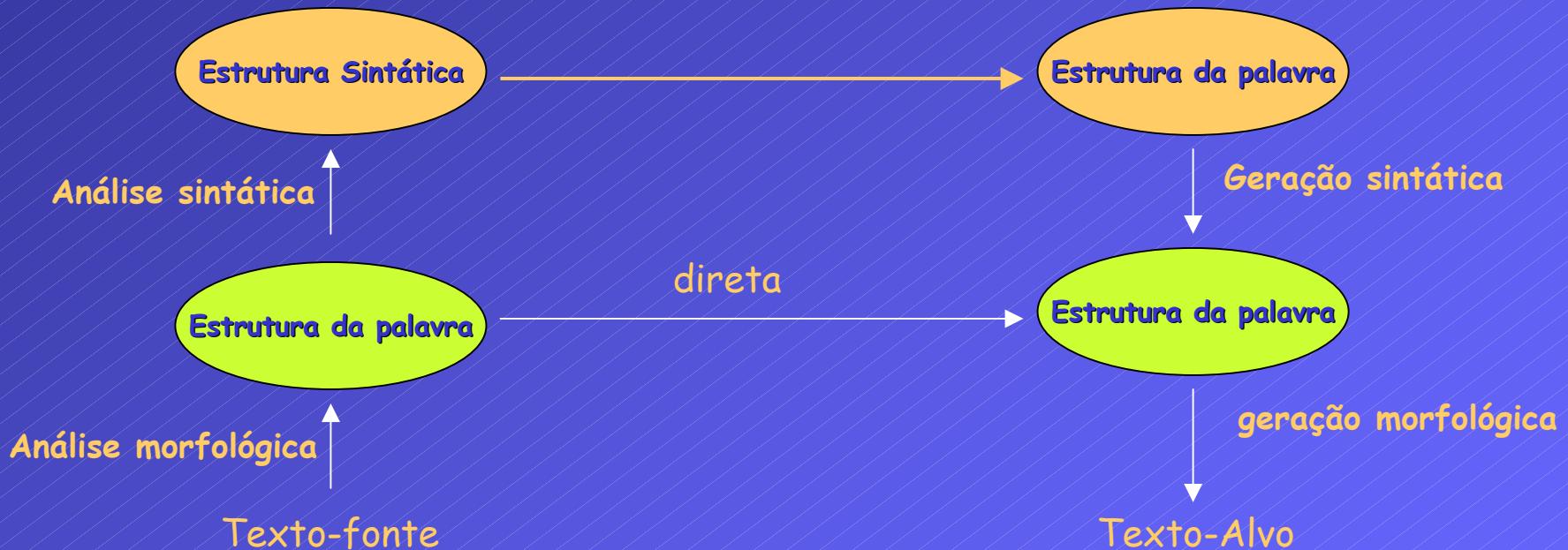
Tradução direta



A Interlíngua UNL

Tradução Sintática

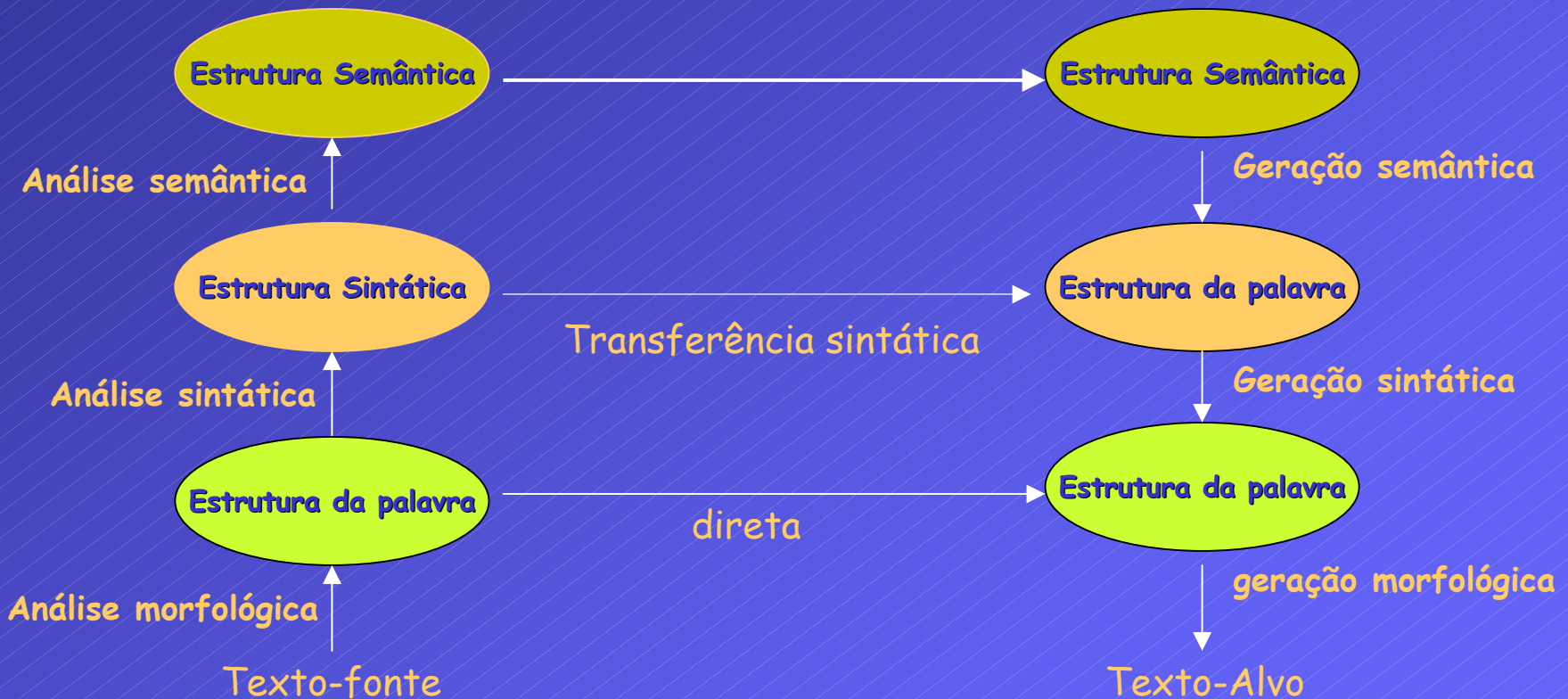
descrever as estruturas sintáticas possíveis ou aceitáveis da língua; ou *decompor* o texto em unidades sintáticas a fim de compreender a maneira pela qual os elementos sintáticos são organizados na sentença.



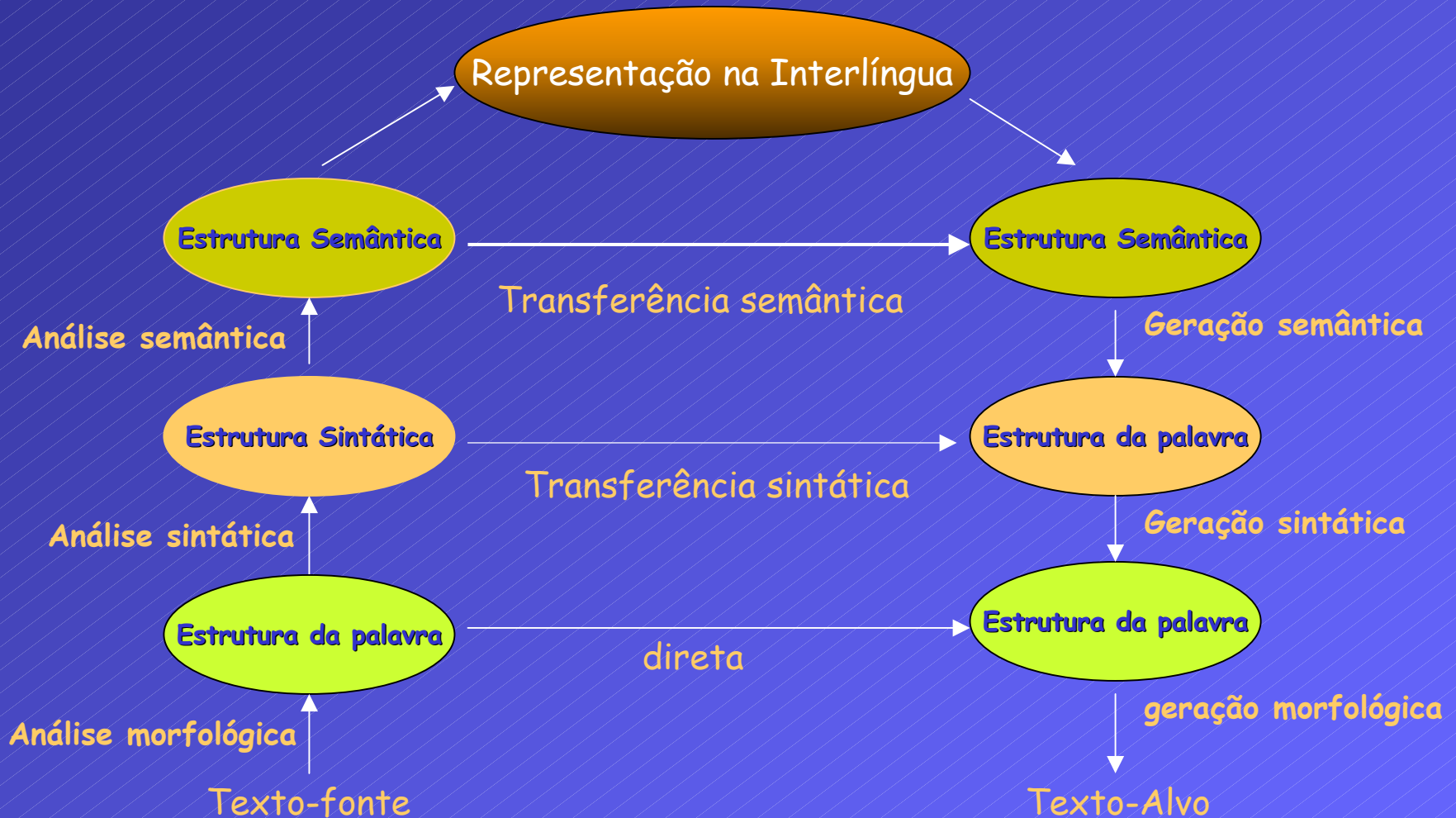
A Interlíngua UNL

Tradução Semântica

Os termos se organizam na oração formando um todo significativo.



A Interlíngua UNL



Blocos UNL ...

- Um conjunto de palavras universais, UWs - *Universal Words*.

"teoricamente trata-se de dicionário universal de conceituação de palavras".

Cada UW expressa um significado único, podendo existir diversas "entradas" para uma mesma palavra em linguagem natural.

... Blocos UNL

- Um conjunto de relações binárias, RLs - Relation Labels.
"teoricamente refere-se a um relacionamento semântico e gramatical universal, entre pares de componentes de sentenças ou palavras universais."
- Um conjunto de atributos, ALs - Attribute Relation.
"atribuir valor gramatical e características pragmáticas de uma palavra universal (UW)."

... Blocos UNL

- As RLs expressam relações semânticas binárias entre UWs contidas numa sentença.
- A representação formal: $RL(UW1,UW2)$.
- 35 RLs (português, 42 geral).
- Exemplo de RL:
 - Agent: `agt(action,thing)`The rabbit runs.
`agt(run.@present,rabbit.@def)`
>> "algo que inicia uma ação".

... Blocos UNL

- Attribute Labels (AL) são usados para especificar relevância gramatical e características pragmáticas de cada componente da sentença.
- Representação formal:
UW.@attrib1.@attrib2...

... Blocos UNL

- AL que especificam um tipo de referência da UW:
 - @generic, @pl, @def, @indef, @not.
- AL que definem os tempos verbais:
 - @past, @present, @future.

... Blocos UNL

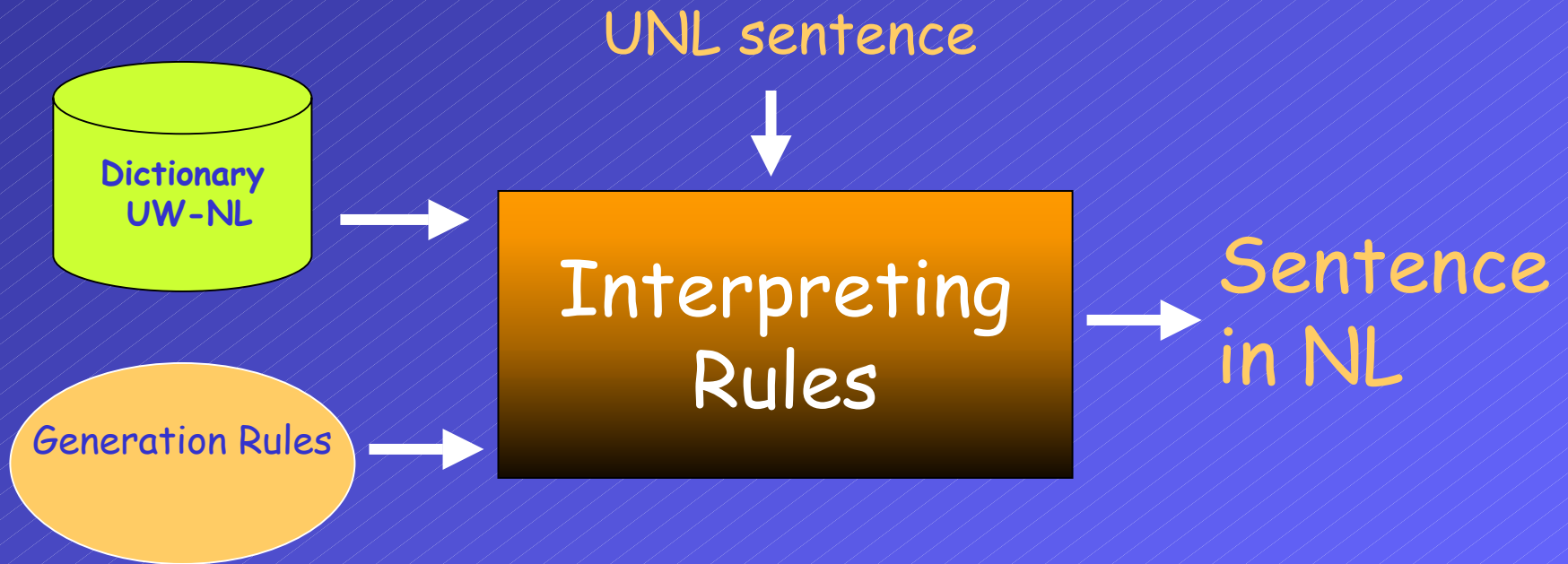
- AL que expressam aspectos:
 - @begin-soon, @begin-just, @end-soon, @end-just, @repeat, @progress.
- AL que expressam o uso das palavras com significado especial para situações particulares (pragmatic information):
 - @focus, @emphasis, @topic, @intention, @recommendation.

... Blocos UNL

```
{unl}
tim(begin(icl>do(obj>thing)).@entry.@past,long ago)
mod(city(icl>region).@def,babylon(icl>country))
plc(begin(icl>do(obj>thing)).@entry.@past,city(icl>region).@def)
agt(begin(icl>do(obj>thing)).@entry.@past,people(icl>person).@def)
obj(begin(icl>do(obj>thing)).@entry.@past,build(icl>do).@past)
agt(build(icl>do).@pred,people(icl>person).@def)
obj(build(icl>do).@pred,tower(icl>building))
aoj(huge(aoj>thing),tower(icl>building))
aoj(seem(aoj>person,obj>thing).@past,tower(icl>building))
obj(seem(aoj>person,obj>thing).@past,reach(icl>do(gol>thing)).@begin-soon)
obj(reach(icl>do(gol>thing)).@begin-soon,tower(icl>building))
gol(reach(icl>do(gol>thing)).@begin-soon,heaven(icl>region).@def.@pl)
{/unl}
```

The Portuguese UNL decoder

DeCo System - DeConverter UNU/IAS/UNL Center



The Portuguese UNL decoder

Processos executados pelo DeCo:

1. Resolver as relações semânticas entre as UWs em notação da UNL, que são vistas como uma *NodeNet*, juntamente resolução dos seus atributos gramaticais.
2. Controlar as janelas da *NodeList* que contém informações a serem processadas para decodificar as regras.

The Portuguese UNL decoder

O DeCo utiliza um "Heardword Dictionary" trabalhando de acordo com um conjunto de regras em conformidade com o idioma da linguagem alvo (neste caso, português).

As regras estabelecem as modificações na *nodelist* visando gerar a sentença em linguagem natural.

UW-Portuguese headwords dictionary

Aproximadamente 63.000 headwords associadas com UWs.

Selecionadas 2.000 palavras em inglês, apontadas no Longman Dictionary que são consideradas as mais representativas dentre as 66.000 entradas do dicionário, sendo satisfatórias para comunicação verbal (segundo os próprios autores do Longman).

UW-Portuguese headwords dictionary

As "*headwords*" foram categorizadas de acordo com suas classes sintáticas, suas características gramaticais e seus atributos semânticos.

Algumas informações semânticas foram incluídas manualmente (1000 hw do UNL Corpus e UN Charter).

Os atributos anexados são parte do conjunto de 63.000 UW utilizados em dicionário eletrônico brasileiro.

UW-Portuguese headwords dictionary

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

Entrada em inglês da UW "threaten" (ameaçar).

UW-Português - exemplo

Smooth(aoj>movement)

[perfeit] {} perfeito

"smooth(aoj>movement)"(stem,plural,larg,rege(de)(em))<P,0,0>;

communication(icl>connection)

[transmissão] {} transmissão

"communication(icl>connection)"(stem,^alomorfe,fem,2arg,
rege(a)(para)(por)(de),deverbais,comum)<P,0,0>;

[transmissõ] {} transmissão

"communication(icl>connection)"
(steam,alomorfe,plural(es),2arg,rege(a)(para)(por)(de),
deverbais)<P,0,0>;

Aplicação de Regras

- Selecionar a UW a ser processada;
- Verificar a entrada no *heardword* para associar ao *node* as informações para a UW: *heardword* e os atributos gramaticais;
- Uma vez que as informações da UW foram recuperadas, o DeCo inicia a pesquisa por regras de geração visando a construção da frase na linguagem natural.
- As regras escolhidas são aquelas que provavelmente permitirão a derivação de sentenças gramaticais na linguagem natural.
- A geração da regra é aplicada apenas no *nó* da *nodelist*.

Aplicação de Regras

- O processo de geração é controlado por duas janelas:
 - **Janela de Geração (G):**
A janela de geração olha para direita e para a esquerda da *nodelist*, especialmente para as características gramaticais de cada *node* da janela.
 - **Janela de Condição (C):**
 - A janela de condição olha para os vizinhos da janela de Geração (G) para verificar o contexto do processo de geração e verificar as características gramaticais dos *nodes* que estão sendo considerados.

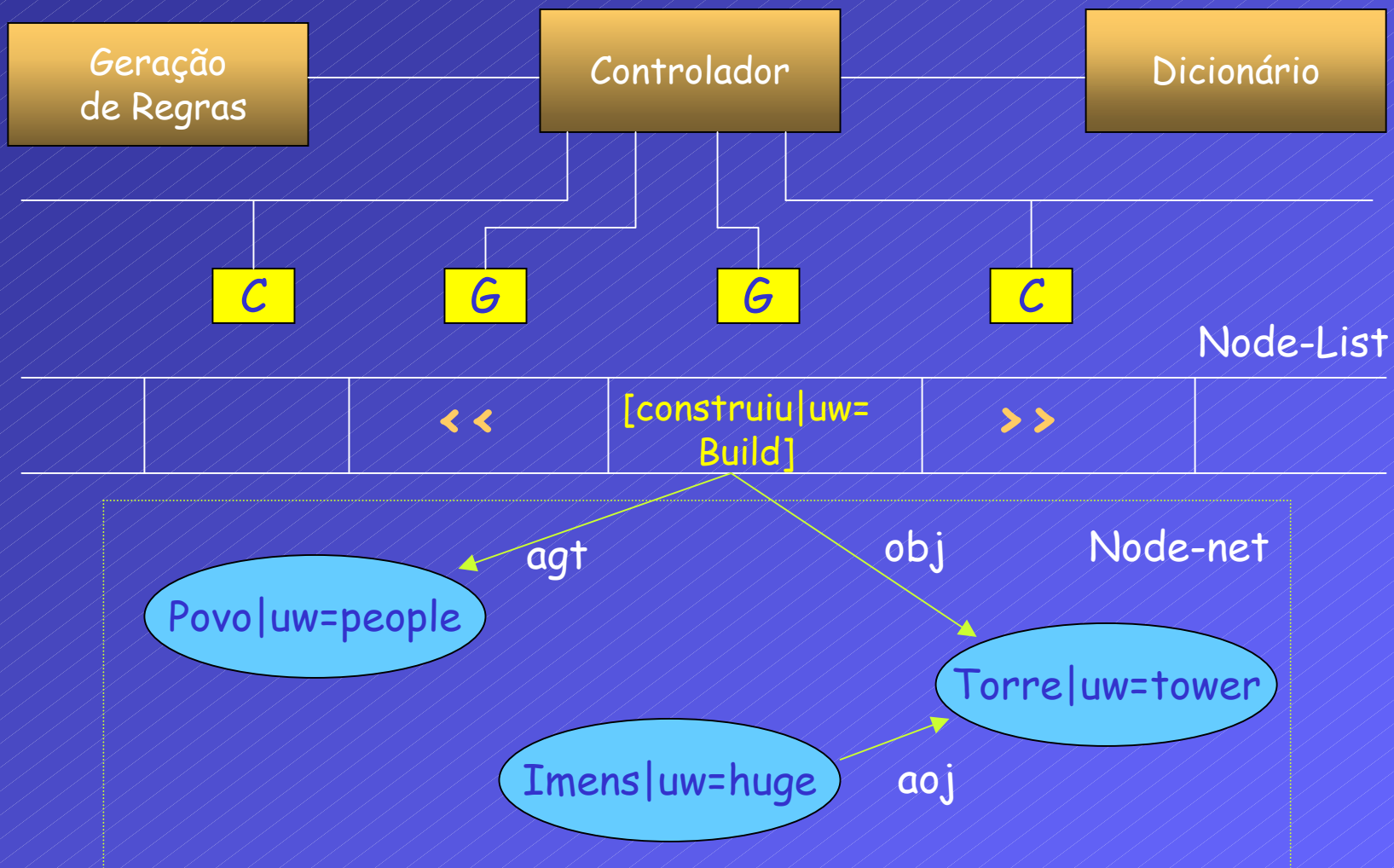
Aplicação de Regras

- Ao verificar os vizinhos dos dois lados da janela de geração(G), a janela de condições(C) ajuda nas futuras modificações da *nodelist*, pois possui informações sobre possíveis aplicações das regras de geração.
- Juntas, as janelas pesquisam informações para gerar regras que melhor combinam com *nodelist* que está sendo processada.

Aplicação de Regras

- A aplicação das regras segue procedendo alterações na *nodelist*:
 - Modificam o *nodelist*: adicionando ou eliminando alguns atributos gramaticais;
 - Inserindo novos *nodes* na posição relativa da janela de geração da *nodelist*.
- Após a inserção, a janela de geração é movimentada.

Aplicação de Regras



Aplicação de Regras

Notação simplificada do exemplo:

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

“O povo constuiu uma torre imensa”

Regras de Geração - Português

Manifestações morfo-sintáticas (UNL):

- Relações semânticas expressas em RLs;
- Atributos gramaticais expressos em ALs;

RLs e ALs são expressas através de construções gramaticais distintas na Língua Portuguesa.

As manifestações gramaticais da notação UNL foram mapeadas para construções lingüísticas do Português, para cada RL e AL.

Regras de Geração - Português

Foram comparadas 20 sentenças UNL do *corpus* da UNU/IAS com o correspondente sentença em português, estilo perfeito.

Foi empregado tradutor de boa qualidade (não apenas um tradutor literal)

RL: expressões gramaticais resultantes de relações semânticas entre componentes das frases e pares de componentes da sentença UNL.

Regras de Geração - Português

AL: expressões gramaticais que contém características morfo-sintática de itens lexicais.

Inadequação Lexical:

"A língua portuguesa comporta um conjunto de locuções e expressões fixas que não admitem variação. Trata-se de expressões cujo sentido deriva, não das partes de que são feitas, mas do todo. Por este motivo, não poderiam sofrer alteração."

Regras de Geração - Português

A adição de itens lexicais permitiu obter a especificação sintática de relações semânticas.

A RL agt foi a manifestação mais freqüente como sujeito da fase

Os resultados com a UNL-português-Brasil obtidos poderiam sugerir que dependem do *corpus* utilizado, podendo não ser representativo da linguagem.

Estudos futuros certamente demandariam um *corpus* mais amplo.

Regras de Geração - Português

Principais manifestações das RLs em português:

RLs	Categorias sintáticas mais freqüentes	Características morfológicas
Soj	Sujeito	Verbo-auxiliar(ou não) substantivo abstrato ou concreto
Obj	Objeto-direto	Verbo agindo como objeto
Agt	Sujeito	Verbo – reforça o substantivo ou pronome pessoal
Tim	Advérbio de tempo	Verbo-advérbio ou expressão adverbial

Regras de Geração - Português

Principais manifestações das RLs em português:

RLs	Categorias sintáticas mais freqüentes	Características morfológicas
Mod	Adjunto nominal e adverbial	Várias classes de palavras
Pla	Advérbio	Verbo-advérbio
Opl	Objeto-direto	Verbo-nominal
Pos	Complemento nominal	Nominal– preposição na frase

Regras de Geração - Português

Principais manifestações das RLs em português:

RLs	Categorias sintáticas mais freqüentes	Características morfológicas
Seq	Coordenadas	Dois verbos em sentenças diferentes
Gol	Objeto direto ou indireto	Sintática não linear
Man	Advérbio	Verbo-advérbio ou expressão adverbial
Ptn	Objeto indireto	Verbo-expressão pré-nominal

Regras de Geração - Português

Principais manifestações das ALs em português:

ALs	Função	Manifestação lingüística
Entry	No principal de um sentença simples ou uma hierarquia entre classes de sentenças	Núcleo do predicado: verbo, núcleo do predicado-verbal ou predicado-nominal em sentenças com o verbo ser. Em sentenças compostas, o verbo expressa conseqüência a respeito do plano antecedente
Present, past, future	Tempo	Predicado verbal ou predicado nominal

Regras de Geração - Português

Principais manifestações das ALs em português:

ALs	Função	Manifestação lingüística
Begin- soom	Evento que acaba de inciar	Advérbio de tempo
Apodosis	Poderia, deveria,??	Oração condicional
State	Estado de um evento finalizado com um resultado	Passado simples

Regras de Geração - Português

Principais manifestações das ALs em português:

ALs	Função	Manifestação lingüística
Progress	Evento em andamento	Expressão verbal: estar mais gerúndio do verbo principal (ndo)
Complete	Evento que já ocorreu	Passado simples
Def	Artigo definido	Artigo definido
Indef	Artigo indefinido	Artigo indefinido

Regras de Geração - Português

Principais manifestações das ALs em português:

ALs	Função	Manifestação lingüística
Pl	Plural	<i>O –s morpheme and its allomorphs</i>
Not	Complemento	Negação de um verbo ou negação de um predicado lexical.

Regras de Geração - Português

As regras de geração que mais ocorreram:

- Inserção à esquerda
:"[o],art,def,masc,sing::"[s,masc,sing,!def:!def:::]P50;
estado inicial do *nodelist*: **menino**
estado final do *nodelist*: **o menino**
- Inserção à direita
:{v,stem,1pes,plural,fut,subj,5,!conjugat:-!conjugat::}
"[armos],dmt,dnp:::"50;
estado inicial do *nodelist*: **cant**
estado final do *nodelist*: **cantarmos**
:{v,vtd,ação,>obj,npred:->obj,+od::}:"s,<obj:-
obj+nod:obj:"P100;
estado inicial do *nodelist*: **constr**
estado final do *nodelist*: **const torre**

Regras de Geração - Português

As regras de geração que mais ocorreram:

- Alteração de atributos
:{suj,masc:::}{adj,psuj,!concorda(gen):-
!concorda(gen),+masc:::}P175
estado inicial do *nodelist*: menino bonit
estado final do *nodelist*: menino binit(>masc)
- Backtracking
?{:::}{plural(alomorfe),@pl:::}^P250
estado inicial do *nodelist*: **intenção**
estado final do *nodelist*: **intenção**

Resultados

Abordagens adotadas nas especificações das regras do DeCo português-brasileiro:

- RLs e ALs foram mapeadas dentro da estrutura morfo-sintática do português-brasileiro;
- Regras morfológicas para palavras inflexionadas foram especificadas, reunido mais de 5000 regras. Um grande número de regras de inserção à esquerda está pronto para a geração de formas inflexionadas dos verbos, sendo 5247 regras de geração para todas as formas verbais. Por outro lado, há apenas uma regra de backtracking, que afeta o desempenho computacional.

Resultados

- Aproximadamente 500 regras estão especificadas para o português brasileiro.

As regras estão assim distribuídas:

Inserções à direita: **0.48%**

Inserções à esquerda: **97.8%**

Alteração de atributos: **1.6%**

Backtracking: **0.01** (apenas uma regra)

Resultados

Ainda que usando um *corpus* limitado, as regras do DeCo, foram significativas para uma considerar uma aplicação mais genérica.

Certamente estas confirmações ocorreram quando o DeCo foi usado para gerar sentenças em português brasileiro do UN Center.

O esforço empregado para realizar este mapeamento (regras gramaticais e atributos semânticos) poderiam ser incorporados ao *Headword Dictionary* português brasileiro.

Resultados

Um trabalho seguinte, seria necessário abordar um grande número de regras morfológicas para tratar todos as possibilidades de tipos de verbos, incluindo os irregulares.

O número de regras poderiam ser reduzidos com novas versões do DeCo, incluindo características que permitissem escolher de forma mais adequada as formas verbais para verbos irregulares.

Exemplos de sentenças

Representação em UNL-BR:

```
obj(function(icl>do).@entry.@obligation,court(icl>judiciary
place):01.@def)
man(function(icl>do)@entry.@obligation,in_accordance_with(icl>manner))
obj(in accordance with(icl>manner), statute(icl>law):01.@def)
aoj(annexed, Statute(icl>law):01.@def)
obj(base(icl>do), statute(icl>law):01.@def)
bas(base(icl>do), statute(icl>law):02.@def)
mod(statute(icl>law):02.@def, court(icl>judiciary place):02.@def)
```

Exemplos de sentenças

...

```
aoj(permanent(icl>state), court(icl>judiciary place):02.@def)
mod(court(icl>judiciary place):02.@def, justice(icl>judiciary))
aoj(international(icl>state), justice(icl>judiciary))
and(form(icl>constitute), base(agt>organization,icl>set,
ppl>place))
obj(form(icl>constitute), statute(icl>law):01.@def)
gol(form(icl>constitute), part(icl>quantity).@indef)
aoj(integral(icl>state), part(icl>quantity).@indef)
mod(part(icl>quantity).@indef, charter(icl>document).@def)
aoj(present(icl>state), charter(icl>document).@def)
```

Exemplos de sentenças

Saida em Portuguese:

"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."

Exemplos de sentenças

O estudo aborda outros exemplos de frases complexas, demonstrando as reais potencialidades da UNL.

Futuras versões do DeCo poderão incorporar regras específicas para melhorar a qualidade das sentenças em português brasileiro.

Seria importante utilizar um *Encoder*, pois as codificações manuais demonstram que podem variar, se montadas por equipes diferentes.

Exemplos de sentenças

Há evidências de que o *Encoder* da UNL tem dependência do idioma.

Dificuldades em alcançar um *Encoder* uniforme também aplica-se à escolha de UWs para representar os conceitos em uma determinada sentença.

Uma saída exata depende de um dicionário que tenha todas as UWs tratadas inteiramente, o que poderia ser alcançado, em breve, com um grande número de UWs.

Exemplos de sentenças

Caso em que a saída do *DeCo*, por restrições do dicionário de UW, reduziu a exatidão da tradução:

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.

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.

Conclusões e trabalhos futuros

A abordagem da UNL traz uma descrição prática de muitos aspectos cruciais do significado das sentenças.

A correspondência morfológica e sintática entre relações semânticas e gramaticais pode ser identificada, permitindo processar a estrutura superficial das sentenças

A UNL apresenta 35 RL que representam o núcleo das relações semânticas entre as UW. (ps: hoje, são 42).

Conclusões e trabalhos futuros

Os autores acreditam que a UNL é suficientemente poderosa para endereçar variações semânticas e lingüísticas em larga escala.

A UNL tenta minimizar a dependência do idioma, permitindo a criação de estruturas textual genéricas para uso na Web.

Por não tratar de uma simples tradução literal, a UNL poderá ser utilizada para desenvolver rapidamente poderosas aplicações

Conclusões e trabalhos futuros

Exemplos boas são ferramentas:

- Codificar e descodificar homepages - sistemas de índices e sumários em várias línguas;
- Indexação automáticos de grandes quantidades do texto em várias línguas.

Conclusões e trabalhos futuros

Além de ampliar os testes com o *DeCo*, o grupo está trabalhando no *Encoder* para o português brasileiro.

O grupo entende que a criação de *Encoder* é uma desafio enorme.