

Tradução Automática com Paráfrases Bilingues Inglês-Português

Anabela Barreiro
barreiro_anabela@hotmail.com

Universidade do Porto & Linguatca
New York University




Simpósio Doutoral da Linguatca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Apresentação

Estrutura

1. Introdução - Tema da Tese e Exemplos
2. Definição de SVC - Verbo de Suporte e Nominalização
3. Importância das SVCs - Observações
4. Paráfrases de SVCs - Exemplos
5. Paráfrases e TA - Problemas com a Tradução de SVCs
6. TA em Foco: Línguas, Recursos e Aplicação - Funcionalidades e Potencialidades do NooJ
7. Desafios Linguísticos: Análise da Língua de Origem - Fenómenos de Cruzamento entre Línguas
8. Dicionários NooJ - Estrutura dos Dicionários da Língua de Origem e da Língua de Chegada - Exemplos
9. Gramáticas NooJ - Estrutura: Metagrafos e Subgrafos - Transferência - Exemplos
10. Capacidades de Tradução do NooJ - Teste a 3 Desafios Linguísticos
11. Questões de Ordem das Palavras - Adjectivo e Nome em SN
12. Representações Específicas das Línguas - Diferenças de Formatos - Datas
13. SVCs - Transferência mantendo o Verbo de Suporte - Verbos Lexicais como Paráfrases
14. Conclusões e Perspectivas
15. Comentários... - e Perguntas?

Simpósio Doutoral da Linguatca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Introdução

Tema

Tema (inicial) da Tese

Análise linguística e formalização computacional de paráfrases curtas bilingues de construções com verbos de suporte (SVCs) para serem utilizadas em tradução automática

Exemplos:

EN *make a call* => PT *fazer um telefonema* = *fazer uma chamada* = *telefonar*

PT *fazer uma investigação* => EN *make an investigation* = *perform an investigation* = *investigate*

Simpósio Doutoral da Linguatca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Definição de SVC

Verbo de Suporte + Nominalização

No âmbito deste trabalho, uma **Construção com Verbo de Suporte (SVC)** é uma construção nominal predicativa constituída por um **Verbo de Suporte** e uma **Nominalização**, onde:

1. o **Verbo de Suporte** é um verbo que tem uma nominalização como um dos seus argumentos (subj., obj., etc.) e que respeita pelo menos um dos seguintes critérios:
 - a) é semanticamente vazio – contribui com pouco ou nenhum significado para a frase, servindo para atribuir os valores de tempo, número e pessoa (ex: *make* em *make an arrangement* ou *fazer* em *fazer uma apresentação*)
 - b) partilha um ou mais argumentos com a nominalização (ex: em *John pays Mary a visit*, *John* é o sujeito de *pay* e de *visit* e *Mary* é o OI de *pay* e o OD de *visit*).
2. a **Nominalização** é que dá o significado à frase; selecciona o seu verbo de suporte e é normalmente o predicado desta

Simpósio Doutoral da Linguatca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Importância das SVCs

Observações

As SVCs são importantes por vários motivos:

- **Transportam unidades de significação**
Assumimos o princípio fundamental de que são as frases elementares e não as palavras isoladas que constituem unidades de significação (o significado não é composicional)
- **São abundantes em muitas línguas**
Os métodos e as conclusões da análise e formalização das SVCs do inglês para o português são aplicáveis a muitas línguas e os recursos resultantes podem ser mapeados entre diferentes pares de línguas
- **O seu estudo é essencial para a sofisticação de sistemas**
Um sistema de tradução automática não pode ignorar o significado de *pay* em *pay attention*, por exemplo. Um tratamento adequado das SVCs pode melhorar muito os resultados no processo de tradução
- **Representam frequentemente alternativas frásicas = paráfrases**
Em quase todos os casos ou na grande maioria, as SVCs podem ser substituídas por outras construções nominais predicativas ou por palavras isoladas simples

Simpósio Doutoral da Linguatca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Paráfrases de SVCs

Exemplos

Várias “boas” alternativas para exprimir **EVENTO [SVC] LUGAR**

Ex:	<table border="0" style="width: 100%;"> <tr> <td style="width: 33%;">"most of the action must"</td> <td style="width: 33%; color: green;">take place</td> <td style="width: 33%;">in the same studio set"</td> </tr> <tr> <td>↕</td> <td></td> <td></td> </tr> <tr> <td>"a maior parte da acção deve"</td> <td style="color: green;">desenrolar-se</td> <td>no mesmo espaço cénico"</td> </tr> </table>	"most of the action must"	take place	in the same studio set"	↕			"a maior parte da acção deve"	desenrolar-se	no mesmo espaço cénico"	
"most of the action must"	take place	in the same studio set"									
↕											
"a maior parte da acção deve"	desenrolar-se	no mesmo espaço cénico"									
•	most of the action must	take place	in the same studio set								
•	most of the action must	occur	in the same studio set								
•	most of the action must	happen	in the same studio set								
•	a maior parte da acção deve	desenrolar-se	no mesmo espaço cénico								
•	a maior parte da acção deve	passar-se	no mesmo espaço cénico								
•	a maior parte da acção deve	realizar-se	no mesmo espaço cénico								
•	a maior parte da acção deve	ter lugar	no mesmo espaço cénico								
•	a maior parte da acção deve	acontecer	no mesmo espaço cénico								
•	a maior parte da acção deve	dar-se	no mesmo espaço cénico								

Simpósio Doutoral da Linguatca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Definição de Paráfrase de SVCs

Paráfrase (no âmbito deste trabalho)

- Técnica de substituição para exprimir e preservar tanto o significado como a gramaticalidade da frase ou construção original
- Troca por uma frase, construção ou palavra semanticamente equivalente
- Mantem uma equivalência conceptual aproximada (*approximate conceptual equivalence*) e não está limitada apenas a relações de sinonímia (cf. Barzilay & McKeown, 2001)

Simpósio Doutoral da Linguateca
Anabela Barreiro Lisboa, 3-4 de Outubro de 2006

Importância das Paráfrases Observações

O estudo das paráfrases desempenha um papel importantíssimo em todas as aplicações de PLN

Para a TA, a paráfrase é um recurso muito valioso.
De que forma podemos beneficiar deste recurso?

- **Desenvolvimento de Sistemas de TA**
o reconhecimento de paráfrases bilíngues como unidades, permite a um sistema de TA com aquisição automática de regras construir regras de transferência para pares alternativos de SVCs ou colocações.
- **Avaliação Automática de TA**
um estudo formal das paráfrases bilíngues é benéfica para a verificação da qualidade da tradução em geral e pode ajudar a resolver problemas relacionados com o facto de existirem formas alternativas e igualmente válidas de traduzir um texto (Pang et al., 2003)

Simpósio Doutoral da Linguateca
Anabela Barreiro Lisboa, 3-4 de Outubro de 2006

Paráfrases e TA Problemas com a Tradução de SVCs

English sentence: *Then the operatives are more likely to make mistakes when set-ups are constantly changing.*

Manual translation from Compara: *E os mecânicos estão mais sujeitos a cometerem erros quando os sistemas estão constantemente a mudar, e isso conduz a um aumento de desperdício.*

MT outputs:

Translation Engine	Result
FreeTranslation	Então o operativo são mais possível cometer erros quando organizações constantemente mudam.
Systran	Então os operatives são mais prováveis fazer erros quando as instalações estão mudando constantemente.
E-T Server	Então os operadores são mais prováveis para tornar erros quando as situações estão mudando constantemente.
Amikai	Então as operações são mais provável cometer erros quando jogos constantemente está mudando.

Tradução errada

Simpósio Doutoral da Linguateca
Anabela Barreiro Lisboa, 3-4 de Outubro de 2006

TA em Foco Línguas, Recursos e Aplicação

- **Pares de línguas**
Inglês-português (principalmente) e português-inglês – eventualmente outras línguas românicas
- **Recursos**
Conversão dos recursos de OpenLogos
Dicionário multilíngue de +100.000 formas canónicas (i.e., não flexionadas) (inglês para as línguas românicas e alemão); paradigmas flexionais para todas as palavras de todas as línguas; informação semântica, aproveitamento de algumas regras sintáctico-semânticas.
Dicionário de +5.000 nominalizações (NOMLEX) e dicionário de +300 verbos de suporte da NYU
- **Aplicação**
Análise e formalização em NooL

Simpósio Doutoral da Linguateca
Anabela Barreiro Lisboa, 3-4 de Outubro de 2006

TA em Foco Funcionalidades e Potencialidades do NooL

- **Novas funcionalidades**
Ainda poucos resultados quantitativos em termos de tradução automática, mas muitos resultados em termos de dados – conversão será “relativamente” rápida
- **Aspectos linguísticos interessantes da tradução**
Alguns resultados qualitativos e grande potencialidade para desenvolvimento a curto prazo

Simpósio Doutoral da Linguateca
Anabela Barreiro Lisboa, 3-4 de Outubro de 2006

Desafios Linguísticos Análise da Língua de Origem e Fenómenos Contrastivos entre Línguas

Até frases muito simples enfrentam desafios como:

- 1- Análise da Língua de Origem
 - Ambiguidade de POS e Homografia
- 2- Fenómenos Contrastivos entre Línguas
 - Análise e Geração
 - Transferência

Simpósio Doutoral da Linguateca
Anabela Barreiro Lisboa, 3-4 de Outubro de 2006

Desafios Linguísticos

Fenómenos Contrastivos entre Línguas

- + Concordância
- + Tempo e Aspecto
- + Preponderância Activa-Passiva
 - Lunches are paid to all employees*
 - > *Pagam-se almoços a todos os empregados*
 - N1 Aux V(Pass) Prep(to) N2 > V(Ref) N1 Prep(a) N2*
- + Ordem das Palavras
 - This is a red table > Isto é uma mesa vermelha*
 - V Det Adj N > V Det N Adj*
 - I gave John a book > Eu dei um livro ao João*
 - V N1 Det N2 > V Det N2 Prep(a) Det N1*
- + Diferentes Padrões para a Representação de Datas
- + SVCs e Paráfrases

Simpósio Doutoral da Linguateca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Dicionários NooJ

Estrutura

Dicionário da Língua de Origem:

- Dicionário NooJ normal com ligação à língua de origem (campos adicionais correspondentes à tradução)
- Lema, POS, informação de flexão e de derivação (FLX mostra como flexionar; DRV mostra como derivar)
- Pode ser bilingue ou multilingue

Dicionário da Língua de Chegada:

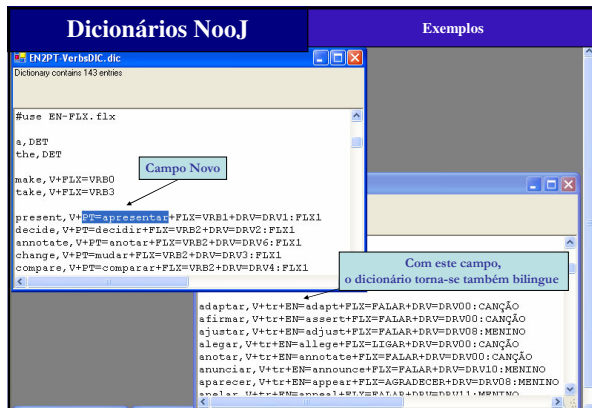
- Tem o mesmo formato que qualquer outro dicionário monolíngue do NooJ, mas é facilmente convertível num dicionário bilingue

Simpósio Doutoral da Linguateca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Dicionários NooJ

Exemplos

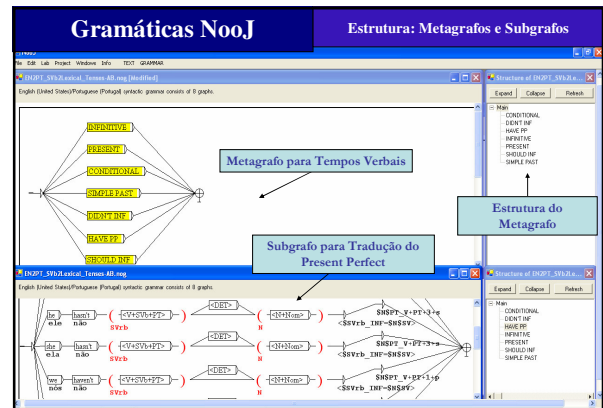


Simpósio Doutoral da Linguateca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Gramáticas NooJ

Estrutura: Metagrafos e Subgrafos

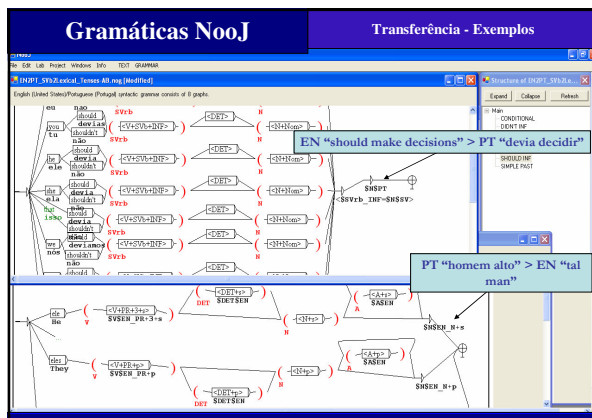


Simpósio Doutoral da Linguateca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Gramáticas NooJ

Transferência - Exemplos



Simpósio Doutoral da Linguateca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Capacidades de Tradução do NooJ

Teste a 3 Desafios Linguísticos

Capacidades de tradução do NooJ em relação aos seguintes aspectos linguísticos:

- Diferenças na ordem das palavras, entre outros
Ex: *homem alto > tall man*
 $N+A > A+N$
- Diferentes representações para datas
Ex: *Monday, September the 11th > segunda-feira, 11 de Setembro*
- SVCs e paráfrases
Ex: *make a visit = visit > fazer uma visita = visitar*

Simpósio Doutoral da Linguateca
Anabela Barreiro

Lisboa, 3-4 de Outubro de 2006

Ordem das Palavras

Adjectivo e Nome no Interior do SN

Dicionário de Português-Ingês

English-Portuguese Dictionary

Portuguese-English Dictionary

Gramática para Ordem de Palavras no Interior do SN

Concordância e Tradução PT-EN

Agreement and Translation PT-EN

Frases de teste

Test Sentences

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

Ordem das Palavras

Adjectivo e Nome no Interior do SN

Dicionário de Português-Ingês

English-Portuguese Dictionary

Portuguese-English Dictionary

Concordância e Tradução PT-EN

Agreement and Translation PT-EN

Gramática para Ordem de Palavras no Interior do SN

Frases de teste

Test Sentences

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

Representações Específicas das Línguas

Diferenças de Formato

As línguas usam formas diferentes de representação de:

- Datas
- Numéricos
- Moradas
- Etc.

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

Representações Específicas das Línguas

Datas

Frases de Teste

Test Sentences

Concordância

Agreement

Dicionário Bilingue

Bilingual Dictionary

Gramática Local

Local Grammar

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

SVCs

Transfêrência como Verbos de Suporte

Dicionário de Inglês - Língua de Origem

English-Portuguese Dictionary

Local Grammar

Concordância - Identificação e Tradução

Agreement - Identification and Translation

Texto de Corpora

Corpus Text

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

SVCs

Verbo Lexical como Paráfrase

Textos de Corpora

Corpus Text

Local Grammar

Concordância - Identificação e Tradução

Agreement - Identification and Translation

Texto de Corpora

Corpus Text

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

SVCs Paráfrases – Verbos Lexicais Plenos

The screenshot displays two windows. The top window, titled "Gramática Local", shows a flowchart of verb forms: INFINITIVE to make -> I -> I'm presenting -> I present. The bottom window, titled "Dicionário de Inglês – Língua de Origem", shows a tree structure for the verb "make" with various grammatical forms like "make", "making", "made", "makes", "made", "making", "made", "making".

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

SVCs Paráfrases – Verbos Lexicais Plenos

The screenshot shows a window titled "Concordância – identificação e tradução" with a list of English sentences and their corresponding Portuguese paraphrases. For example, "I had to make up my mind" is translated as "Tive de fazer uma escolha". Below the list is another window titled "Gramática Local" showing a flowchart similar to the first slide.

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

Dicionários NooJ

The screenshot shows the NooJ interface with a list of verbs and nominalizations. A box highlights "NOMLEX" with "Nominalizações +5.000" and "Verbos de Suporte +300".

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

Conclusões e Futuro

Conclusões
Boas perspectivas + trabalho

Futuro
Desenvolvimento de um módulo de tradução automática inglês-português e português-inglês para as paráfrases
Mais otimização e novas funcionalidades
Propagação automática de valores como a concordância
Capacidade de criar gramáticas mais simples e fáceis de construir – especificação do lema
Avaliação baseada em tradutores humanos e automáticos

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006

Comentários...

Perguntas?

São bemvidos!

Obrigada!

Anabela
)

Simpósio Doutoral da Linguística
Anabela Barreiro
Lisboa, 3-4 de Outubro de 2006