

# Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation

**David Vilar, Daniel Stein, Hermann Ney**

**IWSLT 2008, Honolulu, Hawaii  
20. October 2008**

**Human Language Technology and Pattern Recognition  
Lehrstuhl für Informatik 6  
Computer Science Department  
RWTH Aachen University, Germany**

# 1 Introduction

- ▶ **Hierarchical phrase-based models: Generalization of phrase-based-models**
  - ▷ **Allow for “gaps” in the phrases**
  - ▷ **Integration of reordering in the translation model**
- ▶ **Study the effect of extraction heuristics**
- ▶ **Extension with inclusion of (soft) syntactic features**

# Outline

- 1 Introduction**
- 2 Hierarchical Phrases**
- 3 Heuristic Features**
- 4 Syntactical Features**
- 5 Experimental Results**
- 6 Conclusions**

## 2 Hierarchical Phrases

- ▶ Formalization as a synchronous CFG
- ▶ Rules of the form  $X \rightarrow \langle \gamma, \alpha, \sim \rangle$ , where:
  - ▷  $X$  is a non-terminal
  - ▷  $\gamma$  and  $\alpha$  are strings of terminals and non-terminals
  - ▷  $\sim$  is a one-to-one correspondence between the non-terminals of  $\alpha$  and  $\gamma$
- ▶ Example:

$X \rightarrow \langle \text{中 } X^{\sim 0} \text{ 那个 } X^{\sim 1}, \text{It's the } X^{\sim 1} \text{ in the } X^{\sim 0} \rangle$

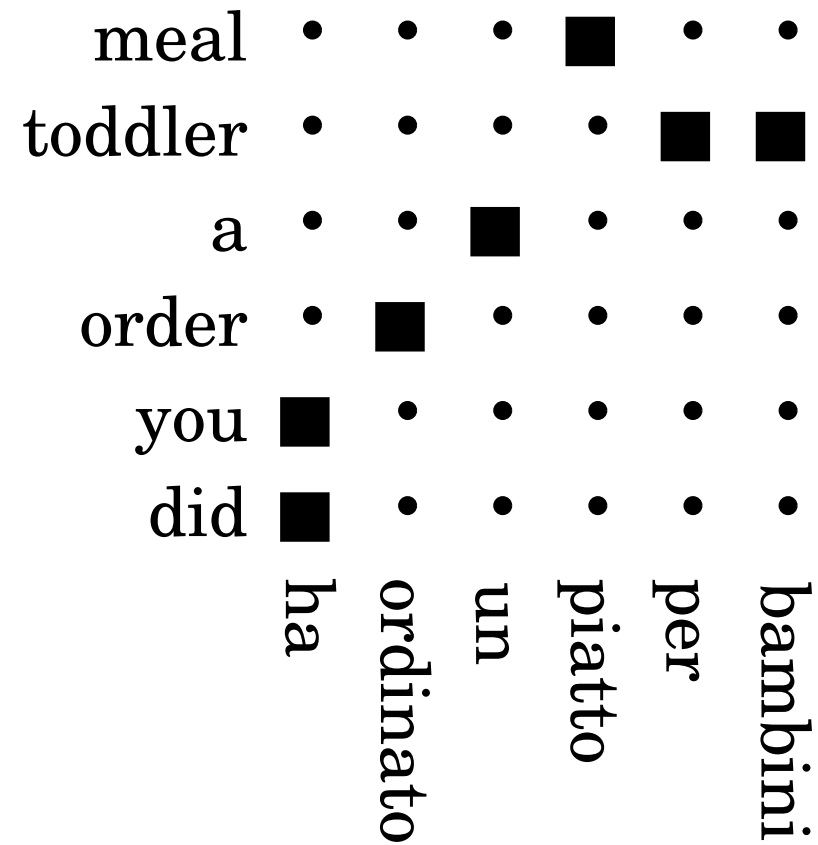
$X \rightarrow \langle \text{也 要 } X^{\sim 0} \text{ 一些 } X^{\sim 1}, \text{like to } X^{\sim 0} \text{ some } X^{\sim 1} \text{ too} \rangle$

- ▶ Additionally: Glue rules

$S \rightarrow \langle S^{\sim 0} X^{\sim 1}, S^{\sim 0} X^{\sim 1} \rangle$

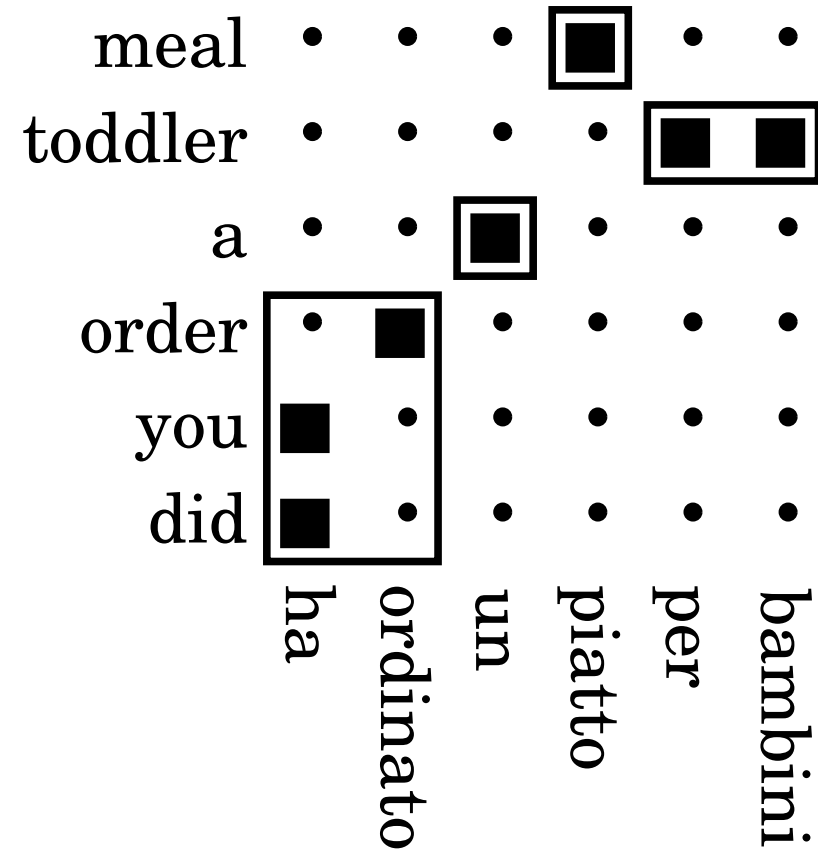
$S \rightarrow \langle X^{\sim 0}, X^{\sim 0} \rangle$

# Illustration



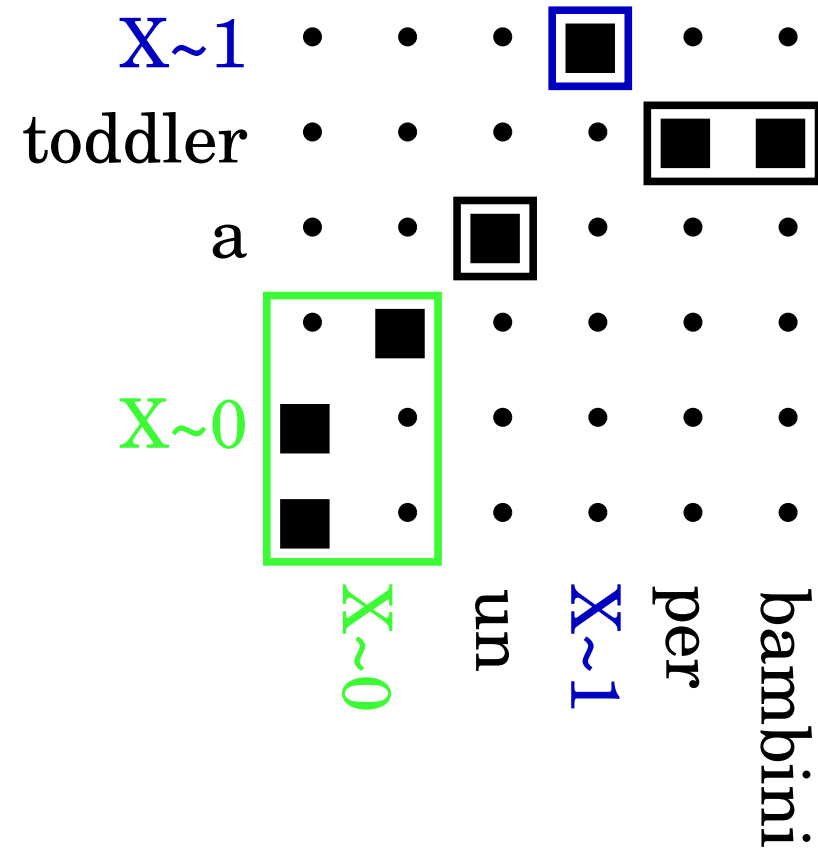
## Alignment

# Illustration



Standard phrases

# Illustration



## Example rule

### 3 Heuristic Features

► Following features were tested:

**Paste rule** Binary feature for rules of the form

$$X \rightarrow \langle X^{\sim 0} \alpha, X^{\sim 0} \beta \rangle \text{ or } X \rightarrow \langle \alpha X^{\sim 0}, \beta X^{\sim 0} \rangle$$

**Hierarchical penalty** Binary feature for hierarchical rules

**Number of non-terminals** Two binary features indicating if the rule has one or two non-terminals.

**Extended glue rule** added rule of the form

$$X \rightarrow \langle X^{\sim 0} X^{\sim 1}, X^{\sim 0} X^{\sim 1} \rangle$$

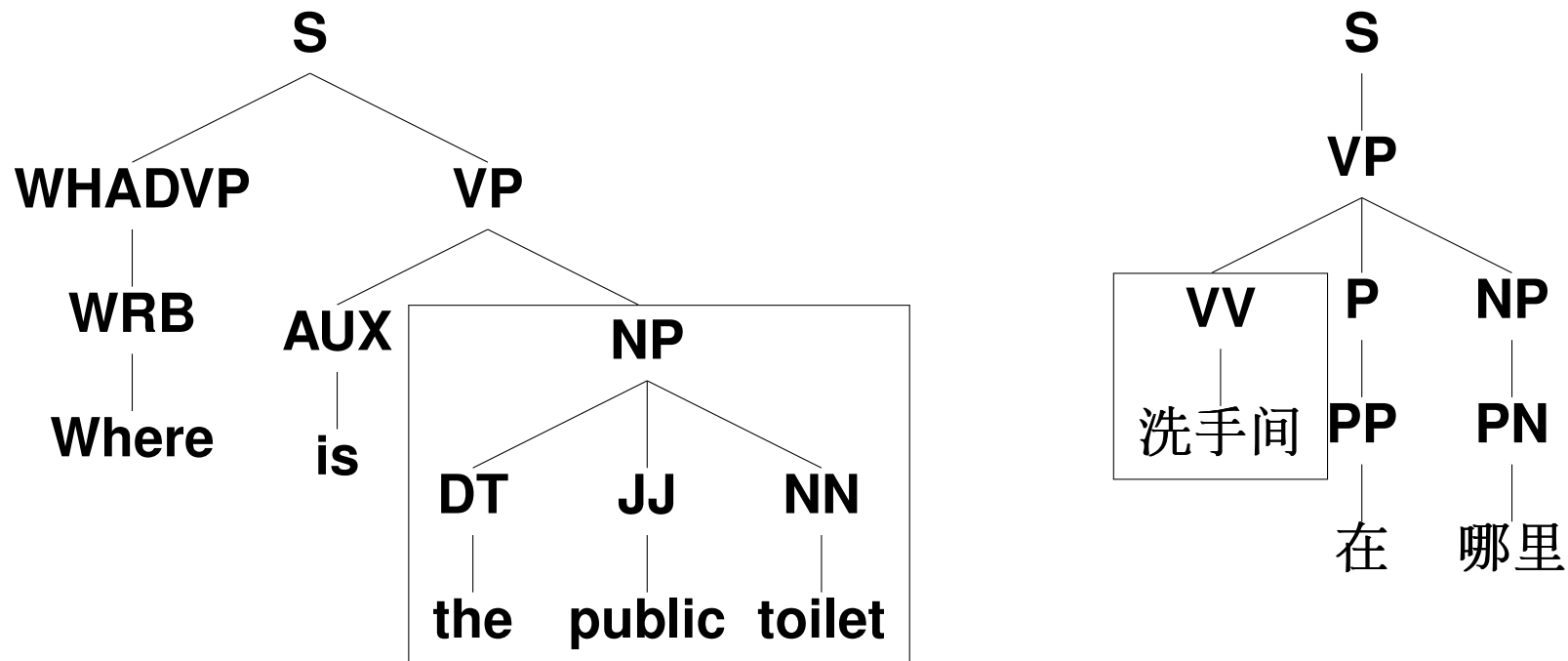


# 4 Syntactical Features

- ▶ **Goal: include linguistic information from a deep syntactic parser**
- ▶ **Idea: introduce additional soft syntactic features**
- ▶ **This can be done during the extraction of the phrases**
  - ▷ **No additional computational costs during decoding**
  - ▷ **Can be done both on source and target side**
  - ▷ **Rules are not filtered out**

# “Valid” syntactical phrases

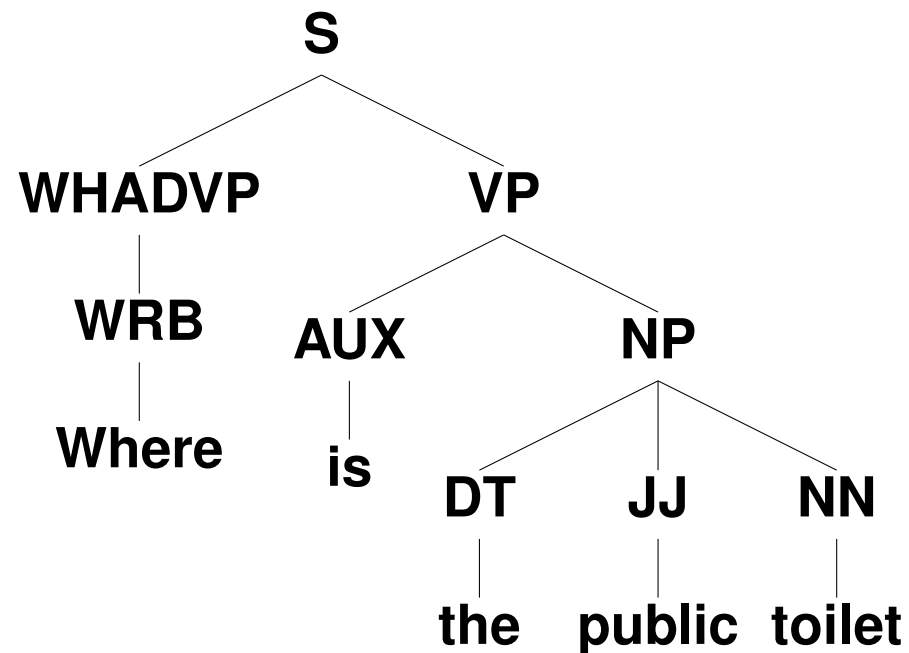
- ▶ A phrase is valid when a node exists that completely covers all positions
- ▶ In order to obtain a normalized score, we add up all the counts and divide by the number of occurrences of the phrase pair



Extracted rule:  $X^{\sim 0}$  在哪里 # Where is  $X^{\sim 0}$

# Scoring variants

$m(i, j)$  = minimum number of words to be deleted or added to a phrase, so that it fits the yield of a node

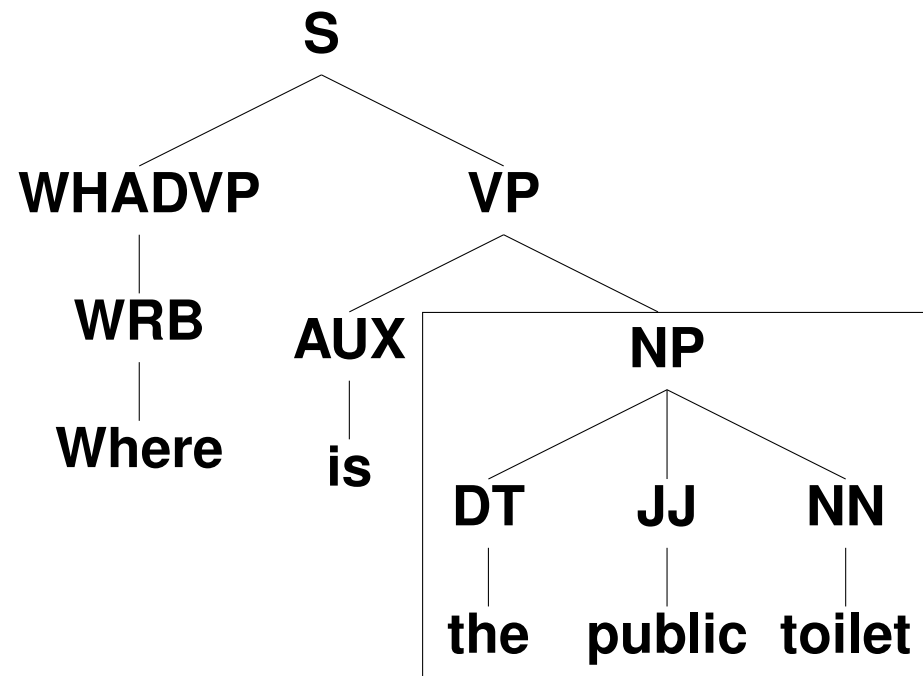


Source Phrases:

- ▶ public toilet
- ▶ is the

# Scoring variants

$m(i, j)$  = minimum number of words to be deleted or added to a phrase, so that it fits the yield of a node

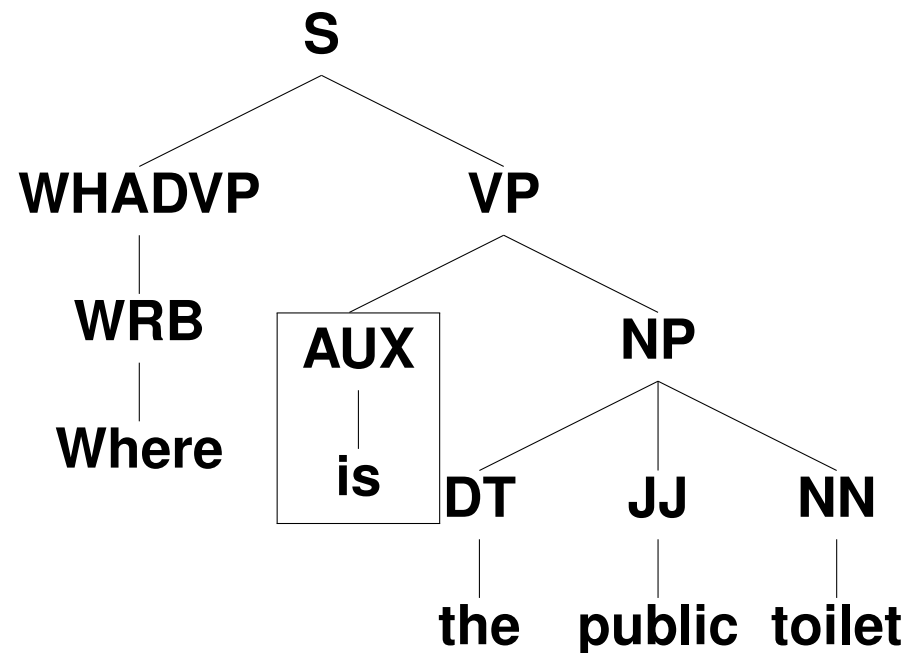


Source Phrases:

- ▶ public toilet  $m(i, j) = 1$
- ▶ is the

# Scoring variants

$m(i, j)$  = minimum number of words to be deleted or added to a phrase, so that it fits the yield of a node



## Source Phrases:

- ▶ **public toilet**  $m(i, j) = 1$
- ▶ **is the**  $m(i, j) = 1$

► Four count (“smoothing”) variants:

$$c(i, j|t) := \begin{cases} \delta(m(i, j), 0) & \text{binary} \\ \frac{1}{m(i, j) + 1} & \text{linear} \\ \frac{1}{\exp(m(i, j))} & \text{exponential} \\ \frac{j - i}{(j - i) + m(i, j)} & \text{relative} \end{cases}$$

# 5 Experimental Results

## ► IWSLT BTEC Data (Tourist and Travel domain)

		Chinese	English
<b>Training data</b>	<b>Sentences</b>	<b>23 940</b>	
	<b>Running words</b>	<b>181 486</b>	<b>232 746</b>
	<b>Vocabulary</b>	<b>9 041</b>	<b>10 350</b>
<b>Test 2004 Data</b>	<b>Sentences</b>	<b>500</b>	
	<b>Running words</b>	<b>7 543</b>	<b>10 718</b>
	<b>OOVs</b>	<b>96</b>	<b>154</b>
<b>Test 2005 Data</b>	<b>Sentences</b>	<b>506</b>	
	<b>Running words</b>	<b>8 052</b>	<b>10 828</b>
	<b>OOVs</b>	<b>101</b>	<b>164</b>
<b>Test 2008 Data</b>	<b>Sentences</b>	<b>507</b>	
	<b>Running words</b>	<b>6325</b>	
	<b>OOVs</b>	<b>87</b>	

# Results

	test04		test05		test08
	BLEU	TER	BLEU	TER	BLEU
<b>baseline</b>	<b>47.3</b>	<b>42.6</b>	<b>50.9</b>	<b>37.6</b>	<b>39.6</b>
<b>non-syntactic information</b>					
<b>hierarch</b>	<b>48.4</b>	<b>41.9</b>	<b>51.4</b>	<b>38.1</b>	<b>39.6</b>
<b>paste</b>	<b>49.1</b>	<b>41.6</b>	<b>51.1</b>	<b>38.0</b>	<b>40.8</b>
<b>glue2</b>	<b>48.2</b>	<b>41.8</b>	<b>51.2</b>	<b>37.6</b>	<b>39.7</b>
<b>1NT2NT</b>	<b>48.4</b>	<b>42.2</b>	<b>51.8</b>	<b>37.2</b>	<b>39.8</b>
<b>syntactic information</b>					
<b>binary</b>	<b>47.8</b>	<b>41.7</b>	<b>51.7</b>	<b>37.5</b>	<b>40.3</b>
<b>linear</b>	<b>47.6</b>	<b>41.9</b>	<b>51.2</b>	<b>37.6</b>	<b>40.6</b>
<b>exponential</b>	<b>47.9</b>	<b>41.7</b>	<b>51.6</b>	<b>37.4</b>	<b>40.3</b>
<b>relative</b>	<b>47.3</b>	<b>42.4</b>	<b>51.5</b>	<b>37.3</b>	<b>40.2</b>



# Results

	test04		test05		test08
	BLEU	TER	BLEU	TER	BLEU
<b>baseline</b>	<b>47.3</b>	<b>42.6</b>	<b>50.9</b>	<b>37.6</b>	<b>39.6</b>
<b>non-syntactic information</b>					
<b>hierarch + paste</b>	<b>48.5</b>	<b>42.0</b>	<b>51.9</b>	<b>37.6</b>	<b>39.6</b>
<b>hierarch + paste + glue2</b>	<b>49.2</b>	<b>42.5</b>	<b>50.8</b>	<b>37.5</b>	<b>39.5</b>
<b>hierarch + paste + glue2 + 1NT2NT</b>	<b>48.6</b>	<b>41.6</b>	<b>51.0</b>	<b>37.9</b>	<b>40.0</b>
<b>combination of both syntactic and non-syntactic information (all features)</b>					
<b>binary</b>	<b>46.9</b>	<b>42.5</b>	<b>50.6</b>	<b>38.4</b>	<b>39.9</b>
<b>linear</b>	<b>48.0</b>	<b>42.3</b>	<b>51.2</b>	<b>38.0</b>	<b>40.5</b>
<b>exponential</b>	<b>47.7</b>	<b>42.3</b>	<b>51.0</b>	<b>38.4</b>	<b>40.0</b>
<b>relative</b>	<b>47.8</b>	<b>42.3</b>	<b>51.0</b>	<b>38.0</b>	<b>40.3</b>

# Example Translations

<b>reference</b>	<b>Where is the exchange counter ?</b>
<b>baseline</b>	<b>The currency exchange office is</b>
<b>syntactical</b>	<b>Where is the currency exchange office ?</b>

<b>reference</b>	<b>Could you exchange it for a new one ?</b>
<b>baseline</b>	<b>You can buy a new one ?</b>
<b>syntactical</b>	<b>Could you change it for a new one ?</b>

<b>reference</b>	<b>You can take our airport shuttle bus to pick up the car .</b>
<b>baseline</b>	<b>You can take our airport shuttle bus with me .</b>
<b>syntactical</b>	<b>You can take our the airport shuttle bus come to pick it up .</b>

# 6 Conclusions

- ▶ **Analyzed heuristics for phrase extraction**
- ▶ **Introduced soft syntactic constraints**
  - ▷ **Use of source- and target-side information**
  - ▷ **No additional search effort**
- ▶ **High variability of results**
  - ▷ **Test on bigger corpora**
- ▶ **Bigger improvements when dealing with speech input (system talk tomorrow!)**
- ▶ **Applicable also to phrase-based systems**

# Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation

**David Vilar, Daniel Stein, Hermann Ney**

**IWSLT 2008, Honolulu, Hawaii  
20. October 2008**

**Human Language Technology and Pattern Recognition  
Lehrstuhl für Informatik 6  
Computer Science Department  
RWTH Aachen University, Germany**

