# RBMT–SMT Hybrid Approach

## Brainstorming Session

**Cristina España** and **Lluís Màrquez**
TALP Research Center
Tecnhical University of Catalonia

MOLTO workshop – GF meets SMT
Göteborg, November 5, 2010

# Talk Overview

# Rule Based MT

- Transfer style translation
- Several sequential steps:
  - → Parse input sentence
  - → Apply structural and lexical transfer rules
  - → Generate output text in the target language
- Transfer grammar: one per language pair
- Parser and generator: one per language

# Rule Based MT

- Transfer style translation
- Several sequential steps:
  - → Parse input sentence
  - → Apply structural and lexical transfer rules
  - → Generate output text in the target language
- Transfer grammar: one per language pair
- Parser and generator: one per language

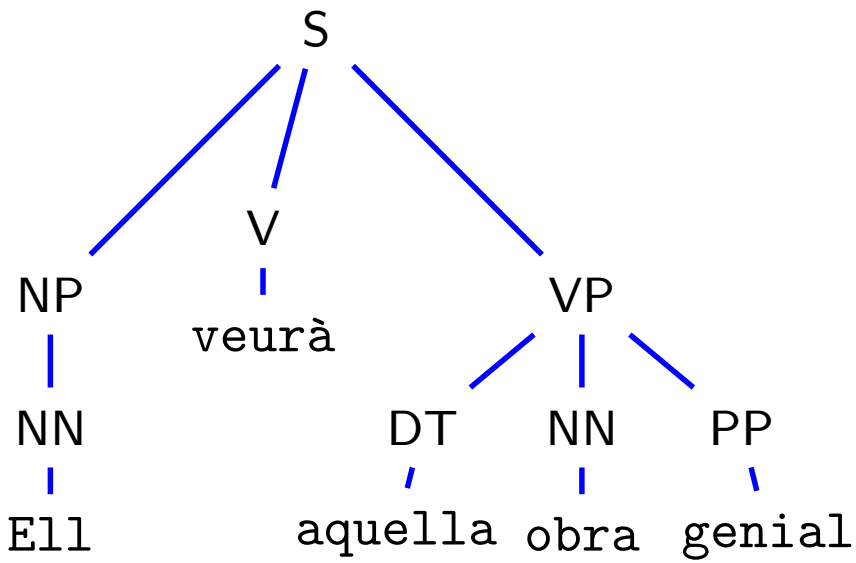
# Rule Based MT

- Transfer style translation
- Several sequential steps:
  - → Parse input sentence
  - → Apply structural and lexical transfer rules
  - → Generate output text in the target language
- Transfer grammar: one per language pair
- Parser and generator: one per language

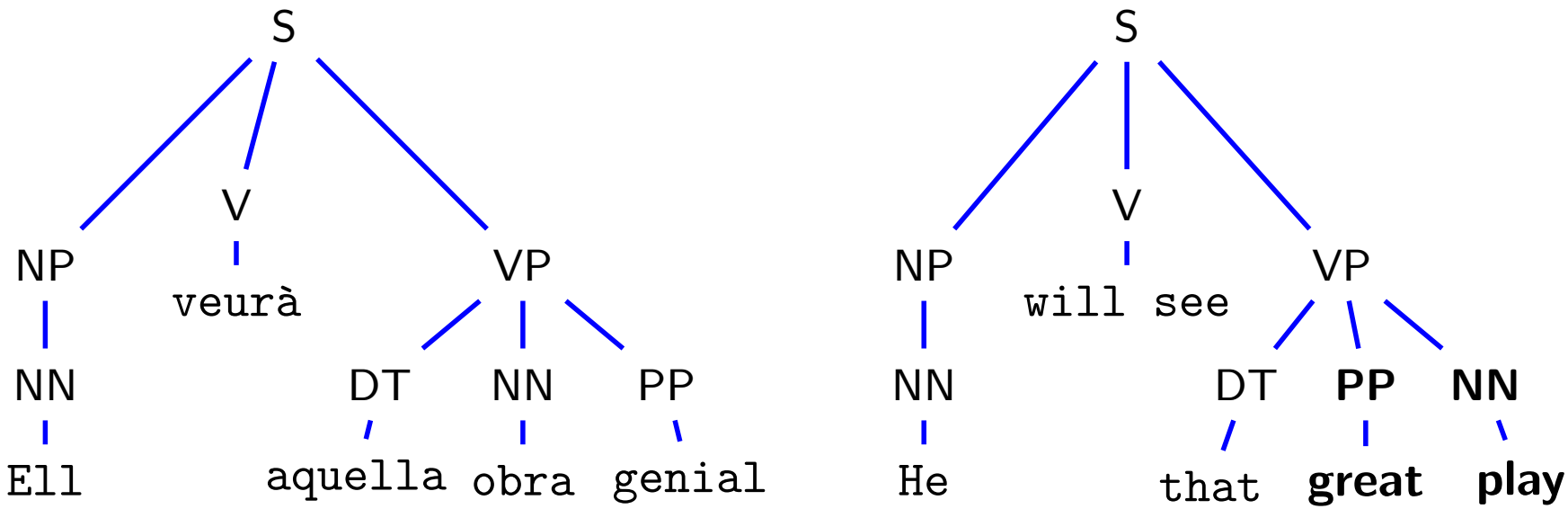
# Rule Based MT: Example

```
Ell veurà aquella obra genial
```

# Rule Based MT: Example

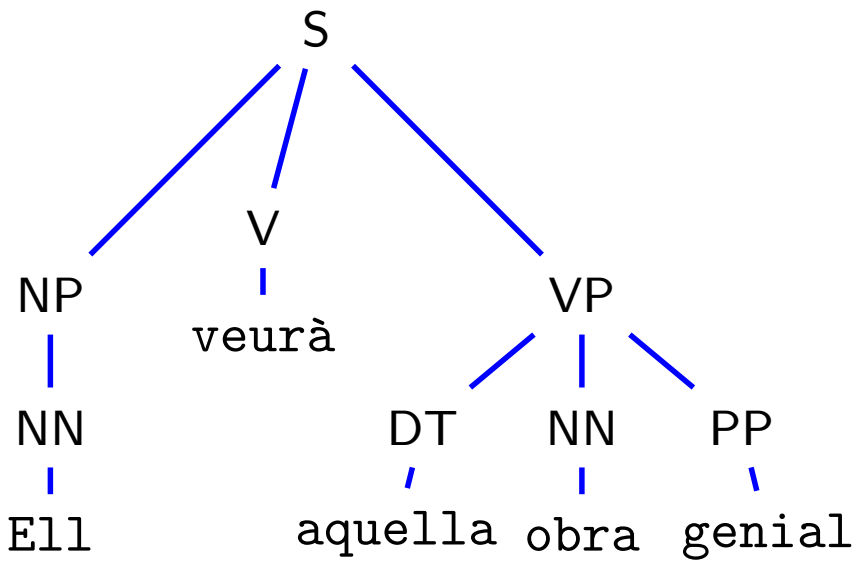


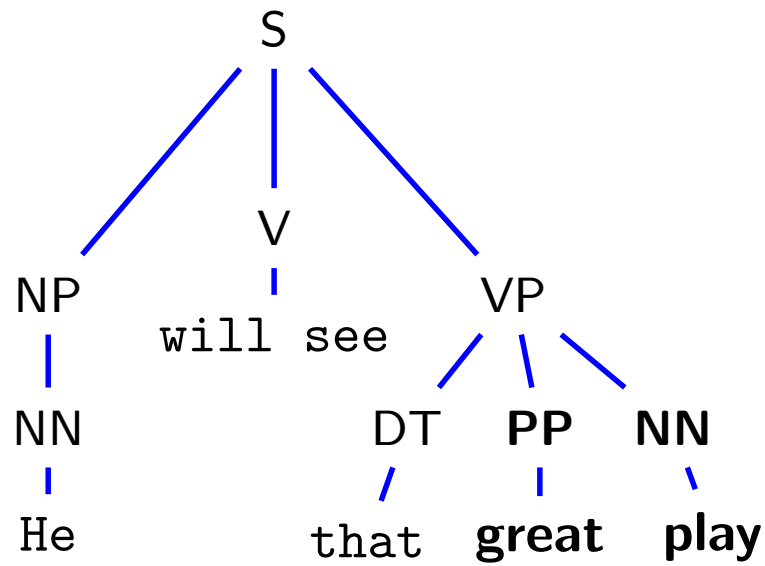
# Rule Based MT: Example




Transfer Rule: $\text{NN} + \text{PP} \Longrightarrow \text{PP} + \text{NN}$

# Rule Based MT: Example




He will see that great play

# Rule Based MT: Pros and Cons

- **Pros** (compared to SMT)
  - → Capture long distance relations and reordering
  - → Better grammaticality
  - → More robust to domain changes
- **Cons**
  - → Dependence on the initial parsing
  - → Lexical transfer disambiguation
  - → High development cost of the grammars and associated resources

# Rule Based MT: Pros and Cons

- **Pros** (compared to SMT)
  - → Capture long distance relations and reordering
  - → Better grammaticality
  - → More robust to domain changes
- **Cons**
  - → Dependence on the initial parsing
  - → Lexical transfer disambiguation
  - → High development cost of the grammars and associated resources

# Rule Based MT: Hybridization with SMT

- **Opportunity**

  → Statistical MT could alleviate some of the RBMT drawbacks

- **Who leads the hybrid model?**

SMT: RBMT is used to enrich the "translation model" of the SMT system (known approach)

RBMT: SMT is used to provide confidence scored translation options to the RBMT target tree (novel)

→ addresses cons number 1 and 2 of previous slide

# Rule Based MT: Hybridization with SMT

- **Opportunity**

  → Statistical MT could alleviate some of the RBMT drawbacks

- **Who leads the hybrid model?**

SMT: RBMT is used to enrich the "translation model" of the SMT system (known approach)

RBMT: SMT is used to provide confidence scored translation options to the RBMT target tree (novel)

→ addresses cons number 1 and 2 of previous slide

# Hybrid RBMT-SMT system

→ Complement with SMT options the RBMT translation structure

- Approach being applied for Basque-to-Spanish with a RBMT system (Matxin)
- OpenMT-2 Spanish Research Project
- UPC+EHU collaboration
- Applicable to MOLTO?

# Hybrid RBMT-SMT system

→ Complement with SMT options the RBMT translation structure

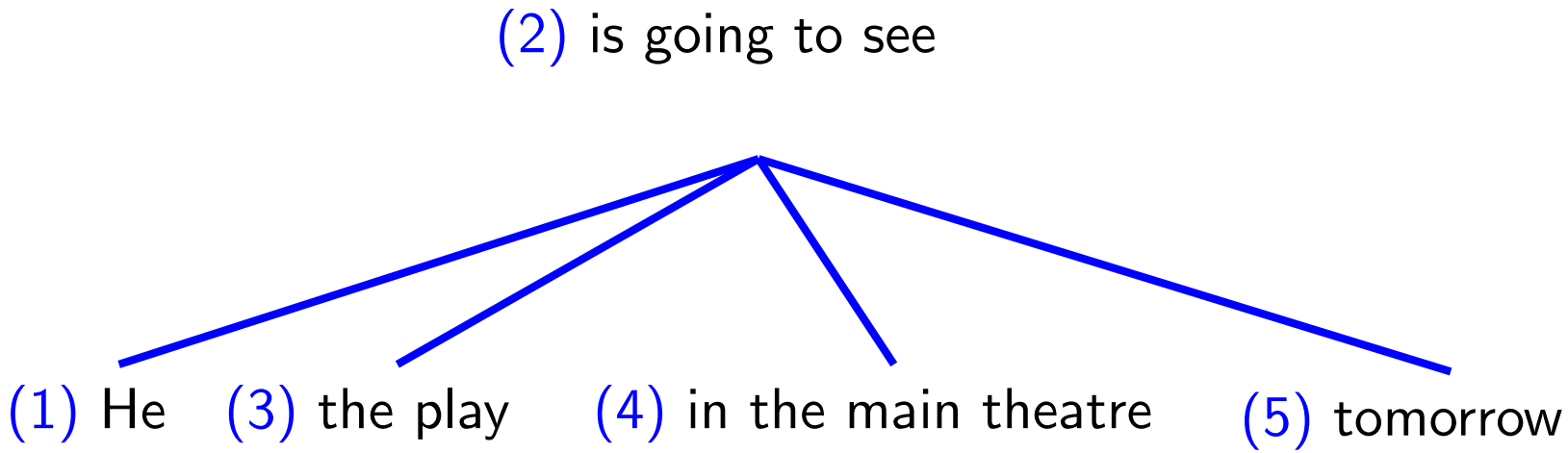
- Approach being applied for Basque-to-Spanish with a RBMT system (Matxin)
- OpenMT-2 Spanish Research Project
- UPC+EHU collaboration
- Applicable to MOLTO?

# Talk Overview

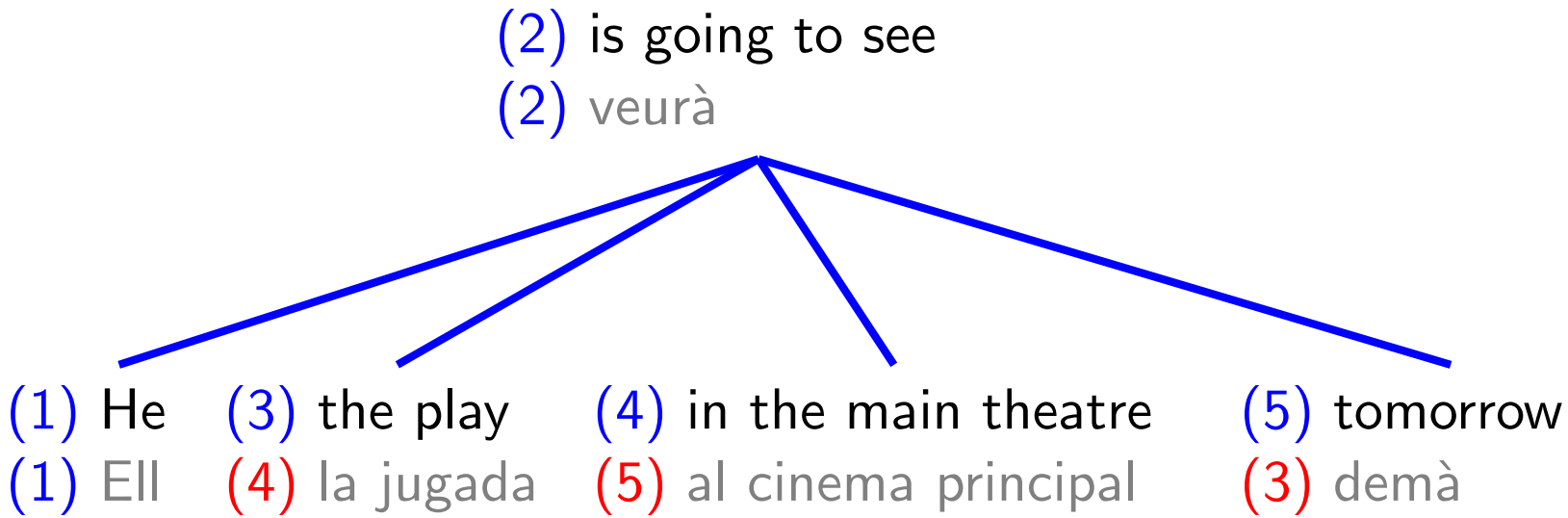
# SMatxinT: Example

He is going to see the play in the main theater tomorrow

# SMatxinT: Example



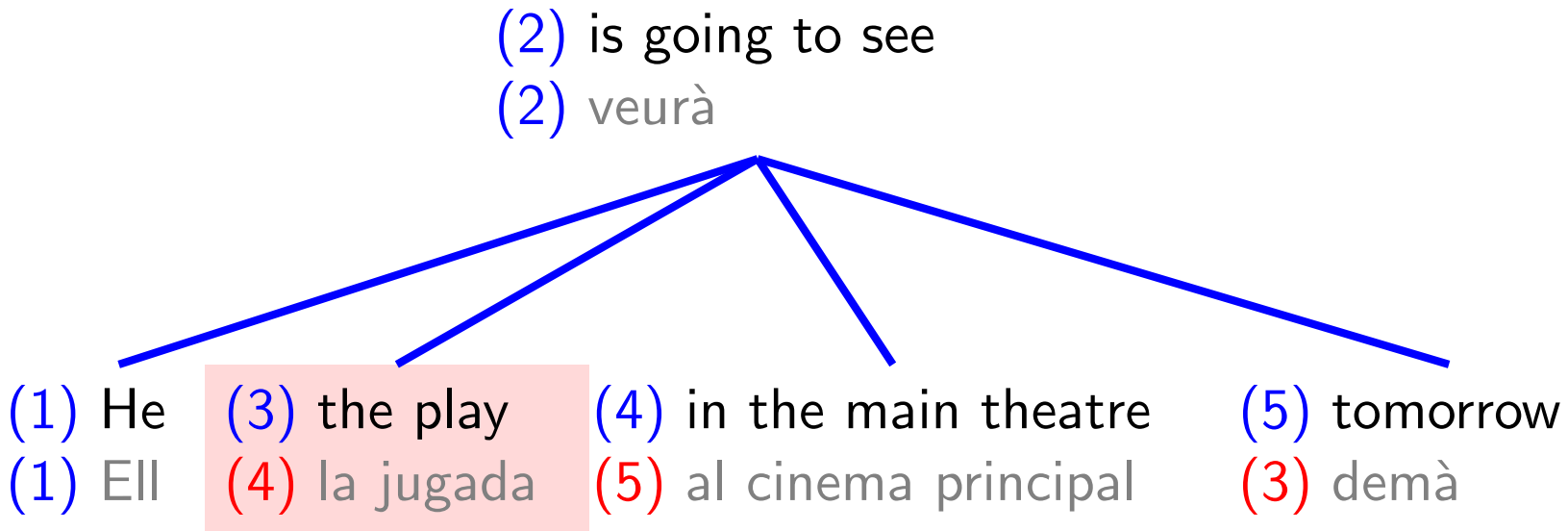
# SMatxinT: Example




SMT: l'obra

SMT: l'obra
l'obra al cinema principal
l'obra al teatre principal
...

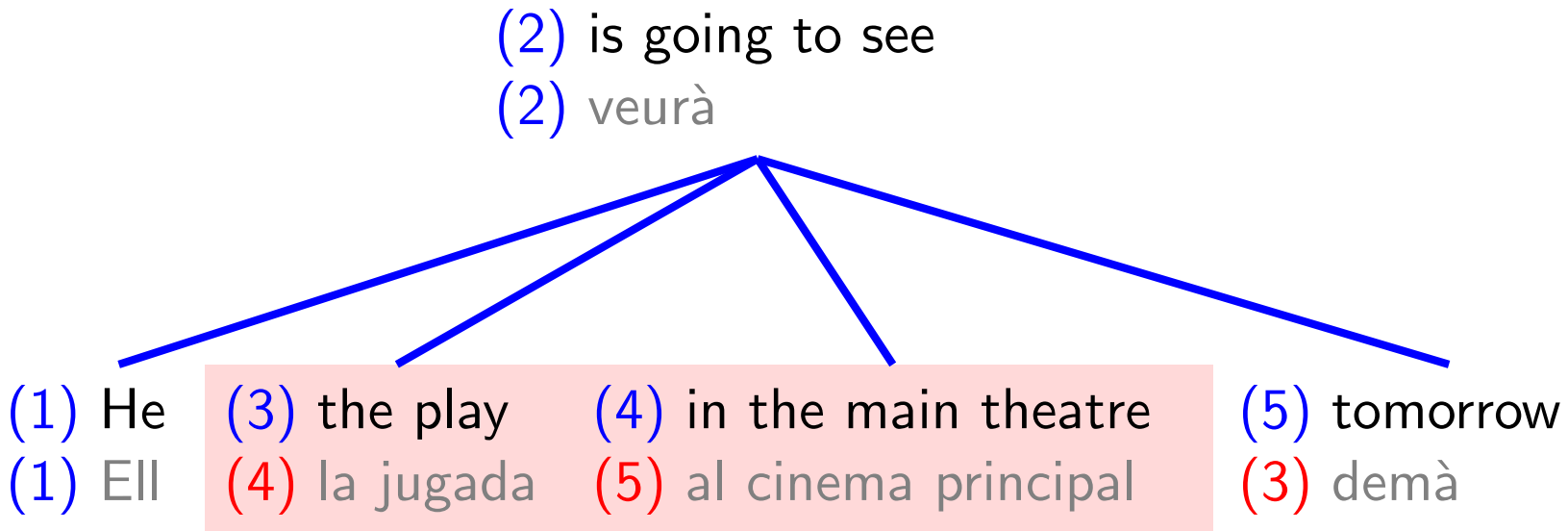
# SMatxinT: Example




SMT: l'obra

SMT: l'obra
l'obra al cinema principal
l'obra al teatre principal
...

# SMatxinT: Example




SMT: l'obra

SMT: l'obra
l'obra al cinema principal
l'obra al teatre principal
...

# SMatxinT: Monotonous Decoding

| (1) | (2) | (5) | (3) | (4) |
|---|---|---|---|---|
| **He** | **is going to see** | **tomorrow** | **the play** | **in the main theatre** |
| Ell | veurà | demà | la jugada | al cinema principal |
| Ell<br>$\phi$ | veurà<br>mirarà<br>... | demà | l'obra<br>la jugada<br>... | al teatre principal<br>al cinema principal<br>al teatre del centre |
| Ell<br>$\phi$ | veurà<br>mirarà<br>... | demà | l'obra al cinema del centre<br>l'obra al teatre principal<br>... | |

...

Anirà a veure demà l'obra al teatre principal
Ell mirarà demà la jugada al teatre principal
...

# SMatxinT: Monotonous Decoding

| (1) | (2) | (5) | (3) | (4) |
|---|---|---|---|---|
| **He** | **is going to see** | **tomorrow** | **the play** | **in the main theatre** |
| Ell | veurà | demà | la jugada | al cinema principal |
| Ell<br>$\phi$ | veurà<br>mirarà<br>... | demà | l'obra<br>la jugada<br>... | al teatre principal<br>al cinema principal<br>al teatre del centre |
| Ell<br>$\phi$ | veurà<br>mirarà<br>... | demà | l'obra al cinema del centre<br>l'obra al teatre principal<br>... | |

...

Anirà a veure demà l'obra al teatre principal
Ell mirarà demà la jugada al teatre principal
...

# SMatxinT: Monotonous Decoding

| (1) | (2) | (5) | (3) | (4) |
|---|---|---|---|---|
| **He** | **is going to see** | **tomorrow** | **the play** | **in the main theatre** |
| Ell | veurà | demà | la jugada | al cinema principal |
| Ell | veurà | demà | l'obra | al teatre principal |
| $\phi$ | mirarà | | la jugada | al cinema principal |
| | ... | | ... | al teatre del centre |
| Ell | veurà | demà | l'obra al cinema del centre | |
| $\phi$ | mirarà | | l'obra al teatre principal | |
| | ... | | ... | |

...

Anirà a veure demà l'obra al teatre principal
Ell mirarà demà la jugada al teatre principal
...

# SMatxinT: Monotonous Decoding

| (1) | (2) | (5) | (3) | (4) |
|---|---|---|---|---|
| **He** | **is going to see** | **tomorrow** | **the play** | **in the main theatre** |
| Ell | veurà | demà | la jugada | al cinema principal |
| Ell<br>$\phi$ | veurà<br>mirarà<br>... | demà | l'obra<br>la jugada<br>... | al teatre principal<br>al cinema principal<br>al teatre del centre |
| Ell<br>$\phi$ | veurà<br>mirarà<br>... | demà | l'obra al cinema del centre<br>l'obra al teatre principal<br>... | |

...

Anirà a veure demà l'obra al teatre principal
Ell mirarà demà la jugada al teatre principal
...

# SMatxinT: Monotonous Decoding

| (1) | (2) | (5) | (3) | (4) |
|---|---|---|---|---|
| **He** | **is going to see** | **tomorrow** | **the play** | **in the main theatre** |
| Ell | veurà | demà | la jugada | al cinema principal |
| Ell<br>$\phi$ | veurà<br>mirarà<br>... | demà | l'obra<br>la jugada<br>... | al teatre principal<br>al cinema principal<br>al teatre del centre |
| Ell<br>$\phi$ | veurà<br>mirarà<br>... | demà | l'obra al cinema del centre<br>l'obra al teatre principal<br>... | |

...

Anirà a veure demà l'obra al teatre principal
Ell mirarà demà la jugada al teatre principal
...

# SMatxinT: Monotonous Decoding

| (1) | (2) | (5) | (3) | (4) |
|---|---|---|---|---|
| **He** | **is going to see** | **tomorrow** | **the play** | **in the main theatre** |

| | | | | |
|---|---|---|---|---|
| **Ell** | veurà | **demà** | la jugada | al cinema principal |
| Ell<br>$\phi$ | veurà<br>**mirarà**<br>... | demà | **l'obra**<br>la jugada<br>... | **al teatre principal**<br>al cinema principal<br>al teatre del centre |
| Ell<br>$\phi$ | veurà<br>mirarà<br>... | demà | l'obra al cinema del centre<br>l'obra al teatre principal<br>... | |

...

Anirà a veure demà l'obra al teatre principal
Ell mirarà demà la jugada al teatre principal
...

# SMatxinT: Summary

- The RBMT system must parse and translate the input sentence
- Phrases and segmentation are those given by the RBMT system
- Each segment (and up) is sent to a generic SMT to provide more partial translations
- A Moses-like decoder is fed with the resulting phrases to search for the highest scored translation
- This statistical decoder performs no reordering and uses very simple features

# SMatxinT: Summary

- Increase robustness
  Usage of multiple input trees by the RBMT

- Select the highest scored translation among the best for each input tree

# Talk Overview

# Preliminary experiment

- **Setting**
  - → Spanish-to-Basque
  - → Matxin and a regular phrase-based SMT system
  - → Training for the SMT system: news + consumer reports + administrative translation memories (<8Mwords)
  - → Matxin: hand developing of the transfer rules; freely available dependency parser for Spanish (FreeLing)
  - → Test set: news domain; 300 Spanish/Basque segments (9150/6343 words); One reference

# Preliminary experiment

- **Setting**
  - → One input tree
  - → No scores in the added partial translations
  - → Only 2 features for the decoder: LM and WP

# Results

| | RBMT | SMT | Hybrid | oracleSBLEU | oracleTER |
|---|---|---|---|---|---|
| BLEU | 4.9 | 7.8 | | | |
| TER | 89 | 86 | | | |

→ Results are not good enough yet (the decoding is extremely simple) but there is a large room for improvement

→ Current work on adding more relevant features and input trees

# Results

| | RBMT | SMT | Hybrid | oracleSBLEU | oracleTER |
|---|---|---|---|---|---|
| BLEU | 4.9 | 7.8 | | **10.5** | 9.6 |
| TER | 89 | 86 | | 81 | **77** |

→ Results are not good enough yet (the decoding is extremely simple) but there is a large room for improvement

→ Current work on adding more relevant features and input trees

# Results

| | RBMT | SMT | Hybrid | oracleSBLEU | oracleTER |
|---|---|---|---|---|---|
| BLEU | 4.9 | 7.8 | 7.7 | **10.5** | 9.6 |
| TER | 89 | 86 | 85 | 81 | **77** |

→ Results are not good enough yet (the decoding is extremely simple) but there is a large room for improvement

→ Current work on adding more relevant features and input trees

# Output analysis

Este partido se disputará a las 19:30 horas en el Fernando Buesa Arena.

RBMT: (0.12, 45.4)
partidu hau 19:30 ordutara eztabaidatuko da Fernando Buesa Arenan .
SMT: (0.18, 63.6)
partida hori da , 19:30ean , fernando buesa arena kiroldegian .
Hybrid: (0.18, 63.6)
partida hori da , 19:30ean , fernando buesa arena kiroldegian .
Oracle: (0.55, 27.2)
partida hori 19:30ean eztabaidatuko da fernando buesa arenan .

Reference: aipatu partida 19:30ean jokatuko da fernando buesa arenan .

# Output analysis

Este partido será dirigido por el trío arbitral, Mitjana, Alzuria y Cardus.

RBMT: (0.19, 54.5)
partidu hau hirukote arbitralak , Mitjana , Alzuria eta Cardus zuzenduko dute .
SMT: (0.31, 100)
partida honetan , trío arbitralak zuzendutako izango da , mitjana , alzuria eta cardus .
Hybrid: (0.31, 100)
partida honetan , trío arbitralak zuzendutako izango da , mitjana , alzuria eta cardus .
Oracle: (0.48, 36.3)
partida hori hirukote arbitralak , mitjana, alzuria eta cardus zuzenduko dute .

Reference: partida hori mitjana , alzuria eta cardus arbitroek zuzenduko dute .

# Output analysis

Con la intención de seguir reforzando el equipo, el Deportivo Alavés ha ofrecido 240.000 euros al Castellón por Dealbert.

RBMT: (0.14, 70.5)
taldea indartu jarraitzeko asmoarekin , Deportivo Alavések 240000 euro eskaini du Castellón Por Dealbertera .
SMT: (0.11, 82.3)
indartu egiten jarraitzeko asmoz , alavesek taldeak eskainitako 240.000 eurokoa da , castelló , dealbert .
Hybrid: (0.13, 70.5)
indartu egiten jarraitzeko asmoz , taldea alavesek 240.000 eurokoa eskaini du castellón por dealbertera .
Oracle: (0.25, 58.8)
taldea indartzen jarraitu asmoarekin , alavesek 240.000 eurokoa eskaini du castellón por dealbertera .

Reference: taldea indartzeko asmoarekin , alavesek 240.000 euro eskaini dizkio castello futbol taldeari dealberten truke .

# Talk Overview

# Application in the MOLTO project

- Complement partial analyses of GF to fill the gaps in cases of unknown words/structures
- Handling of several input analyses (allow for ambiguity)
- **Requirements**
  - → GF has to be previously extended to be robust against unknown words, structures, etc.
  - → A post-processing decoder has to be applied to obtain the best output translation
  - → Statistical parser is needed to help statistical GF on deciding the boundaries of unknown constituents

# Application in the MOLTO project

- Complement partial analyses of GF to fill the gaps in cases of unknown words/structures
- Handling of several input analyses (allow for ambiguity)
- **Requirements**
  - → GF has to be previously extended to be robust against unknown words, structures, etc.
  - → A post-processing decoder has to be applied to obtain the best output translation
  - → Statistical parser is needed to help statistical GF on deciding the boundaries of unknown constituents

# RBMT–SMT Hybrid Approach

## Brainstorming Session

**Cristina España** and **Lluís Màrquez**
TALP Research Center
Tecnhical University of Catalonia

MOLTO workshop – GF meets SMT
Göteborg, November 5, 2010