

# Creación de Recursos Lingüísticos para la Traducción Automática

Victoria Arranz, Núria Castell and Jesús Giménez

TALP Research Center

*III Jornadas en Tecnología del Habla*

Valencia, 17-19 de Noviembre, 2004

# Outline

- **Introduction**
  - **Multilingual Resources and Machine Translation**
- Corpora Developing
- Enriching the Corpora
- Conclusions and Further Work

# Multilingual Resources and Machine Translation

- Statistical Machine Translation (SMT)
  - Domain Oriented  
[Tourist Domain]
  - Syntax-based  
[Johns Hopkins University Summer Workshop, 2003]
- Use of deeper linguistic information  
[ Word -> PoS -> Parsing -> ... ]

# Outline

- Introduction
- **Corpora Developing**
  - **Bilingual Corpora**
  - **Trilingual Corpus**
- Enriching the Corpora
- Conclusions and Further Work

## Bilingual Corpora

- 10,500 short sentences and collocations from
  - recorded speech dialogues
  - touristic websites
  - phrasal books (travel)
- Translation from English into Catalan and Spanish
- Enriched with PoS and lemma  
[Freeling + SVMTool]

# Bilingual Corpora

1613	hello, good morning!	¡hola, buenos días!
1614	oh my goodness !	¡oh Dios mío!
...	...	...
5835	mountain goat	cabra de montaña
5871	a small gratuity	una pequeña propina
...	...	...
10528	additional fee	suplemento
10533	airline company	compañía aérea
10534	airport taxes	tasas de aeropuerto

# Trilingual Corpus

- Spontaneous speech dialogues
- Source
  - Verbmobil database [English] (appointments)
  - Talp-tourism database [Catalan and Spanish] (Accommodation, Flights, Tourist Office)
- Translation [English  $\leftrightarrow$  Catalan and Spanish]

# Trilingual Corpus

00001\_EN: the hotel La\_Habana\_Neptuno in Cuba is located in Havana. it's right on the beach, thirty-five kilometers from the nearest airport.

00001\_ES: el hotel Habana\_Neptuno\_de\_Cuba está situado en La\_Habana. está justo en la playa, a treinta y cinco kilómetros del aeropuerto más próximo.

00001\_CA: l'hotel l'Havana\_Neptuno\_de\_Cuba està situat a L'\_Havana. està just a la platja, a trenta-cinc quilòmetres de l'aeroport més pròxim.

# Trilingual Corpus

	<i>Spanish</i>	<i>Catalan</i>
<i>speech raw time</i>	31h:07m:32s	23h:43m:55s
<i>#speakers</i>	77	56
<i>#dialogues</i>	217	172
<i>#turns</i>	10.998	9.321
<i>#sentences</i>	24.372	19.113
<i>#words</i>	349.970	277.777
<i>#distinct words</i>	11.714	10.057

	<i>Catalan</i>	<i>English</i>	<i>Spanish</i>
Perplexity (3-gram based)	24,9373	19,5766	23,721

# Outline

- Introduction
- Corpora Developing
- **Enriching the Corpora**
  - **Multilingual Resources and XML**
  - **Contents**
  - **Example**
- Conclusions and Further Work

# Multilingual Resources and XML

- ML XML encoding initiatives
  - TEI (CES, XCES)
- Advantages and Drawbacks
  - (+) human/machine-readability, usability, portability, flexibility, robustness...
  - (-) disk/memory space consumption

# Contents

## DOC\_REPOSITORY

- > DOC+
- > SECTION+
- > SEGMENT+
- > LSEGMENT+
- > ACOUSTIC\_EVENTS+
- > WORD+
- > COMPOUND\_WORD+
- > WORD+
-

# Contents

- WORD
  - linguistic features
  - acoustic features
- MULTI-WORDS  
(named entities, compounds, phrasal verbs, light verbs)
- SYNTACTICS (parsing)
- SEMANTICS (role labeling)
- ALIGNMENTS (word-level, phrase-level, concept-level)





## Example 2

• `<BI_DOC_REPOSITORY DATE=' '14/05/2004' ' L1="EN" L2="ES">`

• `<DOC SCN="13" SRC="ES" id="B0078">`

• `<SCT>`

• `<SGM>`

• `<L1> <W L=' 'Meliá' ' P=' 'NNP' ' >Meliá</W>`

• `<W L=' ', ' ' P=' ', ' ' >, </W>`

• `<W L=' 'good' ' P=' 'JJ' ' >good</W>`

• `<W L=' 'morning' ' P=' 'NN' ' >morning</W>`

• `<W L=' '! ' ' P=' '. ' ' >!</W> </L1> ...`

•

## Example 2

...

```

. <L2> <W L='Meliá' P='NP'>Meliá</W>
. <W L=',' P=','>,</W>
. <W L=';' P='Faa'>!</W>
. <W L='bueno' P='AQ'>buenos</W>
. <W L='día' P='NC'>días</W>
. <W L='!' P='Fat'>!</W> </L2>
. </SGM> ...
. </SCT> ...
. </DOC SCN="13" SRC="ES" id="B0078"> ...
. </BI_DOC_REPOSITORY DATE='14/05/2004' L1="EN" L2="ES">

```

# Outline

- Introduction
- Corpora Developing
- Enriching the Corpora
- **Conclusions and Further Work**
  - **Benefits for Machine Translation**
  - **Conclusions**
  - **Further Work**

# Benefits for Machine Translation

- A convenient representation of rich linguistic annotation:
  - morphology
  - syntax
  - named entities
  - semantic roles
  - word alignment
  - ...

# Conclusions

- Use of XML
- DTD for bilingual/multilingual corpora
- Creation of bilingual parallel phrasal lexica
- Creation of a trilingual parallel corpus

## Further Work

- Moving from DTD to W3C XML Schema
- Better Representation of Linguistic Phenomena
  - Syntactic/Semantic Parsing, Multi-words, ...
  - Alignment at different levels (word/phrase)

**Thank you!**

<http://www.lc-star.com>