



Innovación y Conocimiento en la Sociedad Digital

# EDICIÓN 8ª Internet Global Congress

Barcelona, 29 de mayo - 1 de junio, 2006

PALACIO DE CONGRESOS, FIRA BARCELONA, PZA. DE ESPAÑA

La Investigación y el  
 Desarrollo en Tecnologías de  
 Traducción Automática

Rafael E. Banchs, UPC

Conferencia



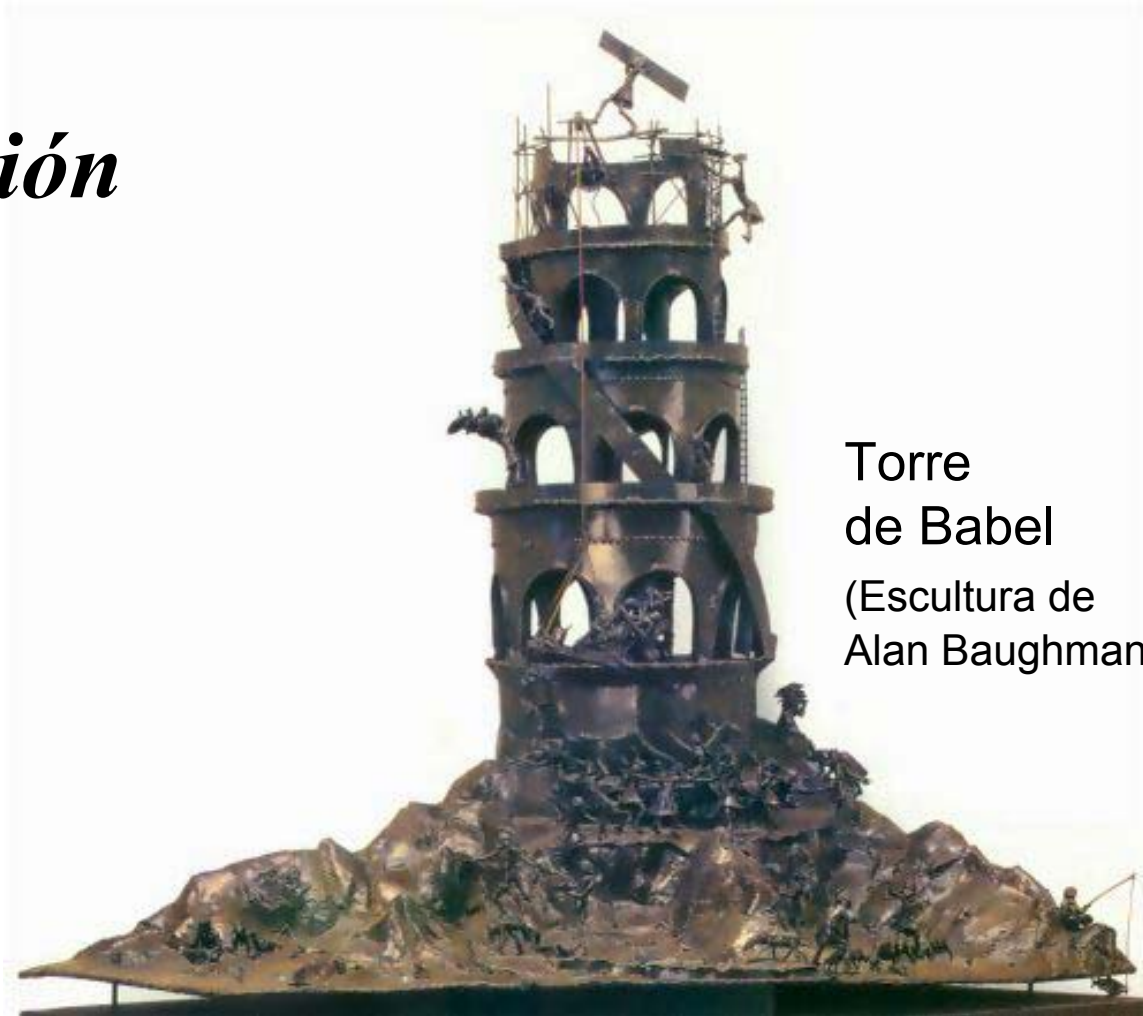
[www.igcweb.net](http://www.igcweb.net)

UN PROYECTO DE:



Fundació  
 Barcelona  
 Digital

# *Introducción*



Torre  
de Babel  
(Escultura de  
Alan Baughman)

## ***La traducción automática como área de investigación***

Búsqueda en <i>www.google.es</i>	<i>Resultados</i>
“ machine translation ”	<b>559.000</b>
“ machine translation ” + research	<b>196.000</b>
“ machine translation ” + research + university	<b>131.000</b>
“ machine translation ” + research – university	<b>63.800</b>
“ machine translation ” + confenerce	<b>123.000</b>
“ machine translation ” + journal	<b>98.100</b>

## ***La traducción automática en el contexto europeo***

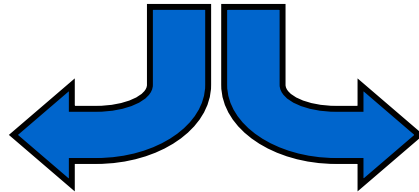
- C-STAR: <http://www.c-star.org>
- EUTRANS: <http://www.cordis.lu/espirit/src/30268.htm>
- VERBMOBIL: <http://verbmobil.dfki.de/verbmobil/overview-us.html>
- LC-STAR: <http://www.lc-star.com>
- NESPOLE!: <http://nespole.itc.it>
- FAME: <http://isl.ira.uka.de/fame/orga.html>
- TC-STAR: <http://www.tc-star.org>

## *Dos paradigmas, cinco métodos*

### Métodos de Traducción Automática

#### Basados en Conocimiento

- Interlingua
  - Transfer
- Traducción directa



#### Basados en Datos

- Traducción basada en ejemplos
- Traducción estadística

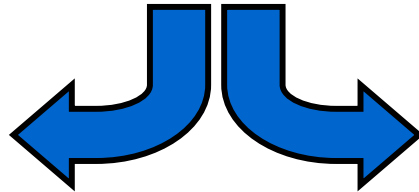


## *Dos paradigmas, cinco métodos*

### Métodos de Traducción Automática

#### Basados en Conocimiento

- Interlingua
  - Transfer
- Traducción directa



#### Basados en Datos

- Traducción basada en ejemplos
- Traducción estadística

# *La aproximación estadística*

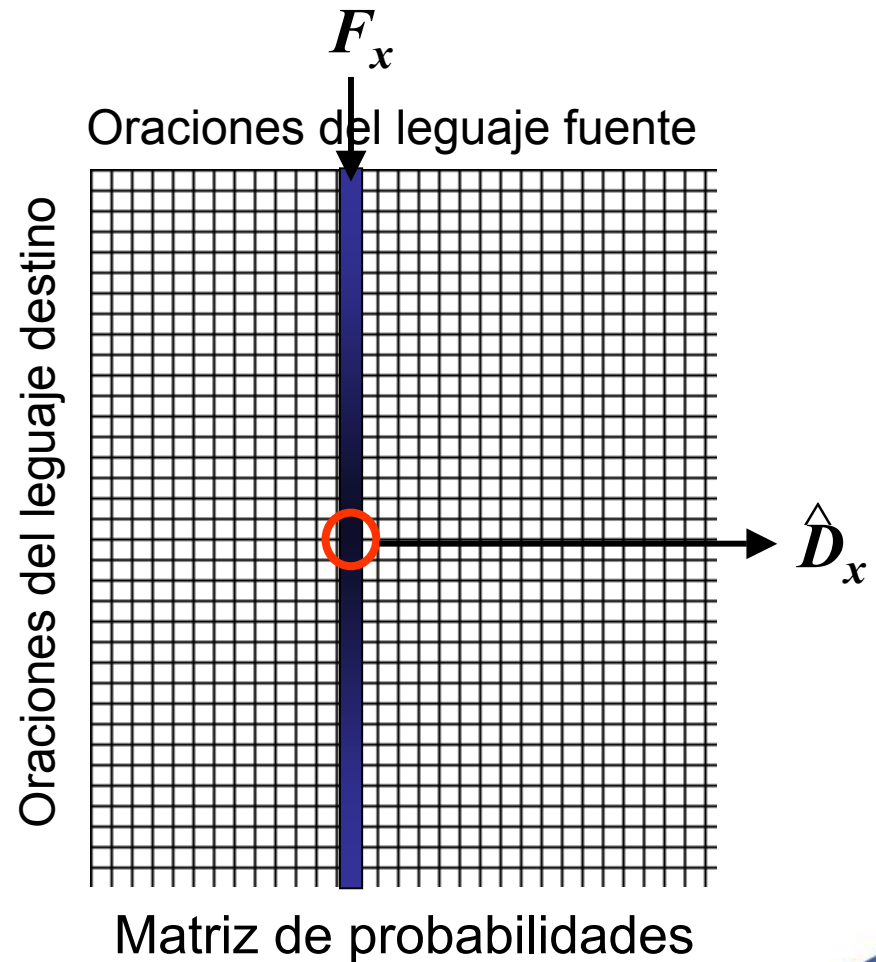


## Planteamiento teórico de la aproximación estadística

$$\hat{D}_x = \underset{D}{\operatorname{argmax}} P(D|F_x)$$

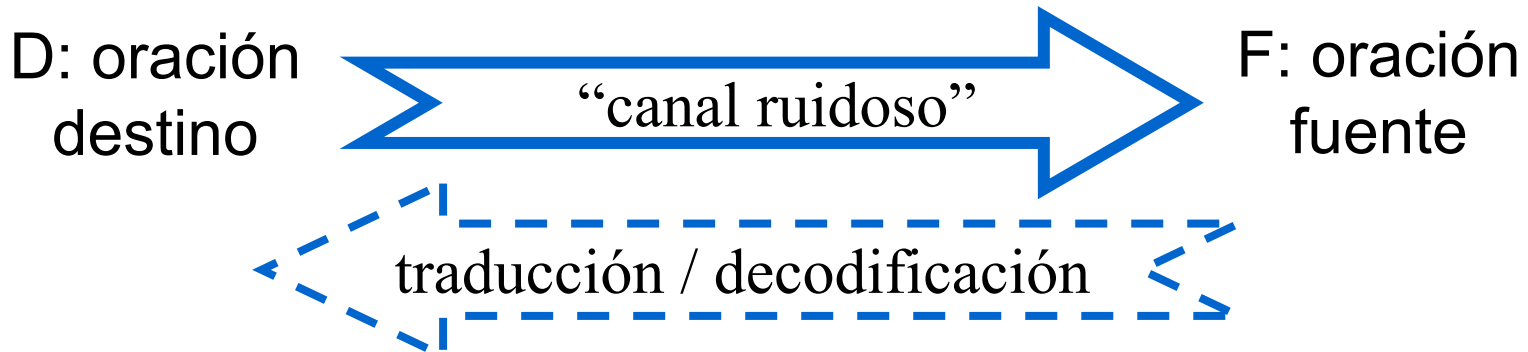
Problemas prácticos:

- Cálculo de las probabilidades
- Espacio de búsqueda





## Primer modelo de traducción estadística



$$\hat{D} = \underset{D}{\operatorname{argmax}} P(D|F) = \underset{D}{\operatorname{argmax}} P(F|D) P(D)$$

Modelo de lenguaje

Modelo de traducción

## ***Implicaciones del modelo***

La búsqueda de la mejor traducción  $P(D|F)$  se convierte en la optimización simultánea de dos características:

- 1.- “*Adequacy*”: búsqueda de los contenidos más adecuados de acuerdo con el modelo de traducción  $P(F|D)$
- 2.- “*Fluency*”: búsqueda de la mejor construcción gramatical de acuerdo con el modelo de lenguaje  $P(D)$

## ***Ejemplo con modelo de lenguaje basado en n-gramas***

<b>Oración</b>	<b>Probabilidad*</b>
“the welcome mr. to sesion president”	<b><math>4,52 \times 10^{-8}</math> (-16,91)</b>
“mr. president welcome to the sesion”	<b><math>1,11 \times 10^{-6}</math> (-13,71)</b>
“sesion the to president mr. welcome”	<b><math>5,02 \times 10^{-9}</math> (-19,11)</b>
“president the sesion to welcome mr.”	<b><math>6,23 \times 10^{-8}</math> (-16,59)</b>
“sesion president welcome to mr. the”	<b><math>8,96 \times 10^{-9}</math> (-18,53)</b>

*\* Probabilidades calculadas con un modelo de 3-gramas entrenado con datos del Parlamento Europeo.*

## ***Ejemplo con modelo de traducción basado en palabras***

### **Oración fuente**

“el presidente vino al parlamento”

### **Probabilidad\***

### **Oraciones destino**

“the president wine to parliament”

**$7,35 \times 10^{-7}$  (-14,12)**

“the president came to parliament”


**$3,42 \times 10^{-8}$  (-17,19)**


“the parliament came to president”

**$3,42 \times 10^{-8}$  (-17,19)**

*\* Probabilidades calculadas con un modelo léxico basado en palabras entrenado con datos del Parlamento Europeo.*

## ***Estado actual del arte: sólo dos cambios importantes***

Canal Ruidoso  Combinación Log-lineal  
*(Och y Ney, 2002)*

Modelo de traducción  
Basado en palabras  Modelo de traducción  
Basado en “Frases”  
*(Zens et al. 2002, Koehn et al. 2003)*

## Combinación Log-lineal de modelos

Enfoque más general, fundamentado en los principios de entropía máxima (*Berger et al. 1996*)

$$\hat{D} = \underset{D}{\operatorname{argmax}} p(D | F) \approx \underset{D}{\operatorname{argmax}} \prod_i p_i(F, D)^{\lambda_i}$$

Canal Ruidoso → caso particular:

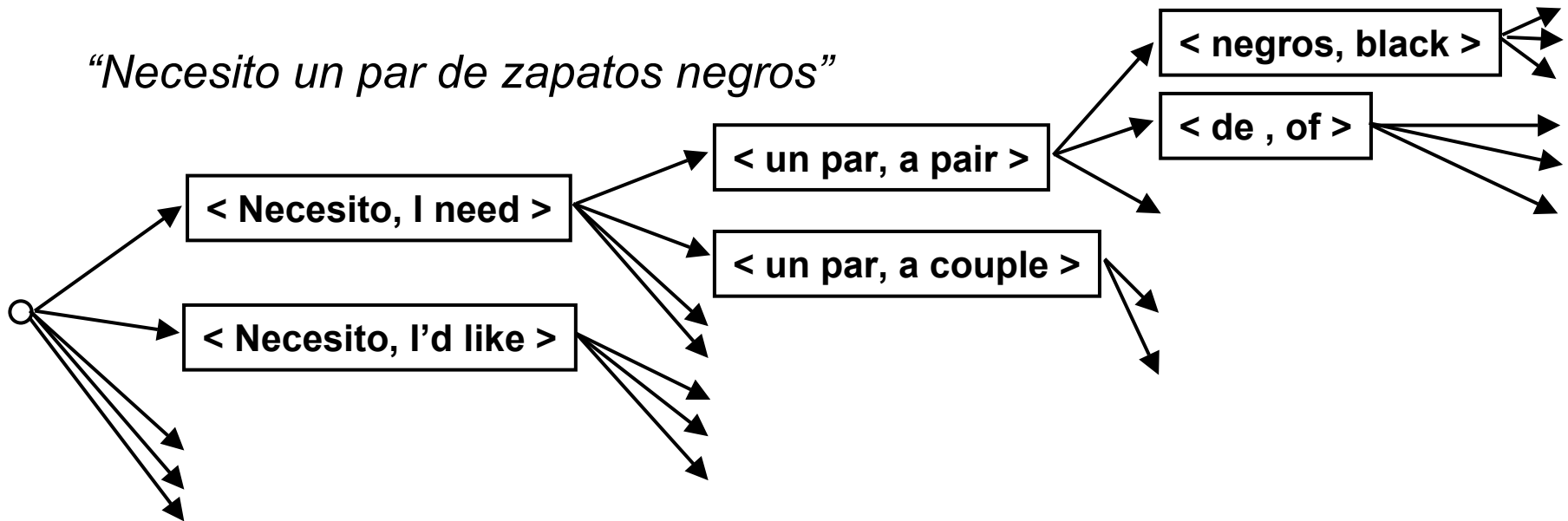
$$p_1(F, D) = p(F|D), \quad p_2(F, D) = p(D), \quad y \quad \lambda_1 = \lambda_2 = 1$$



## Decodificación con modelos basados en “Frases”

Se explora el espacio de las posibles traducciones mediante el uso de un algoritmo de búsqueda (*Wang y Waibel 1997, Tillman et al. 1997, Koehn 2004*)

“Necesito un par de zapatos negros”

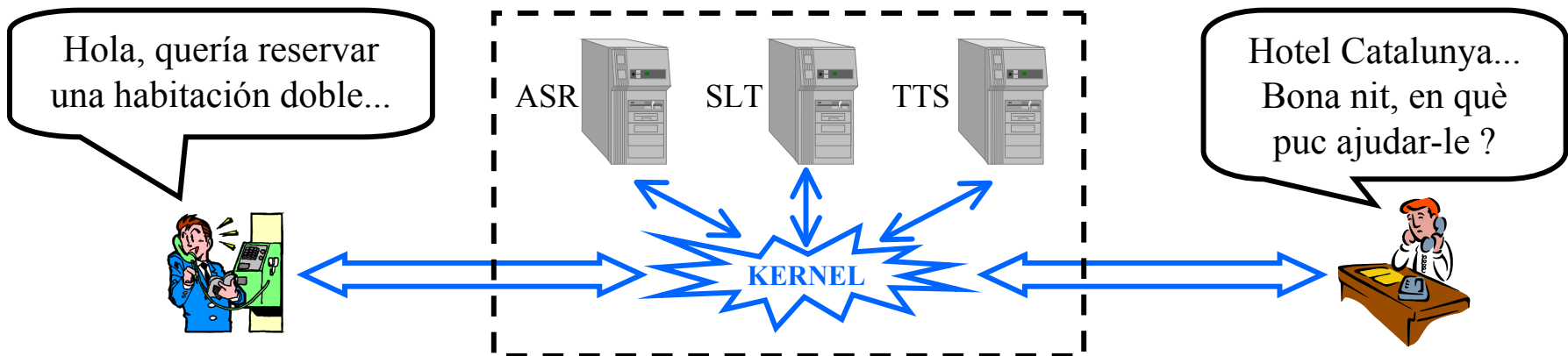


# *Algunos ejemplos experimentales*



## Ejemplo experimental #1: PROYECTO LC-STAR

- <http://www.lc-star.com>
- Prueba de aceptación de una plataforma experimental para comunicación bilingüe entre castellano y catalán
- Datos experimentales y de dominio restringido (turístico)



## ***Descripción de la evaluación***

- 12 participantes para un total de 6 diálogos
- Objetivos de la tarea (reserva de una habitación de hotel):
  - 1.- fecha de llegada
  - 2.- número de noches de la estancia
  - 3.- tipo de habitación requerida
  - 4.- costo por noche del tipo de habitación requerida
  - 5.- nombre completo del cliente
  - 6.- tipo de tarjeta de crédito a ser usada para la reserva
  - 7.- número de la tarjeta de crédito, y
  - 8.- fecha de caducidad de la tarjeta de crédito

## Resultados de la evaluación

Objetivos	dlg1	dlg2	dlg3	dlg4	dlg5	dlg6	obj
día de arribo	0.75	0.60	1.00	-0.50	0.50	-0.63	0.29
noches	1.00	0.38	1.00	1.00	0.50	1.00	0.81
habitación	1.00	1.00	0.50	0.50	0.50	0.43	0.65
precio	-0.50	1.00	1.00	0.60	0.60	0.27	0.50
nombre	–	0.60	1.00	1.00	0.50	0.60	0.74
tipo tc	1.00	1.00	1.00	1.00	0.75	-0.70	0.68
número tc	0.00	-0.25	-0.50	-0.84	-0.57	0.00	-0.36
caducidad tc	1.00	-0.75	0.60	1.00	1.00	-0.75	0.35
<b>diálogo</b>	<b>0.61</b>	<b>0.45</b>	<b>0.70</b>	<b>0.47</b>	<b>0.47</b>	<b>0.03</b>	



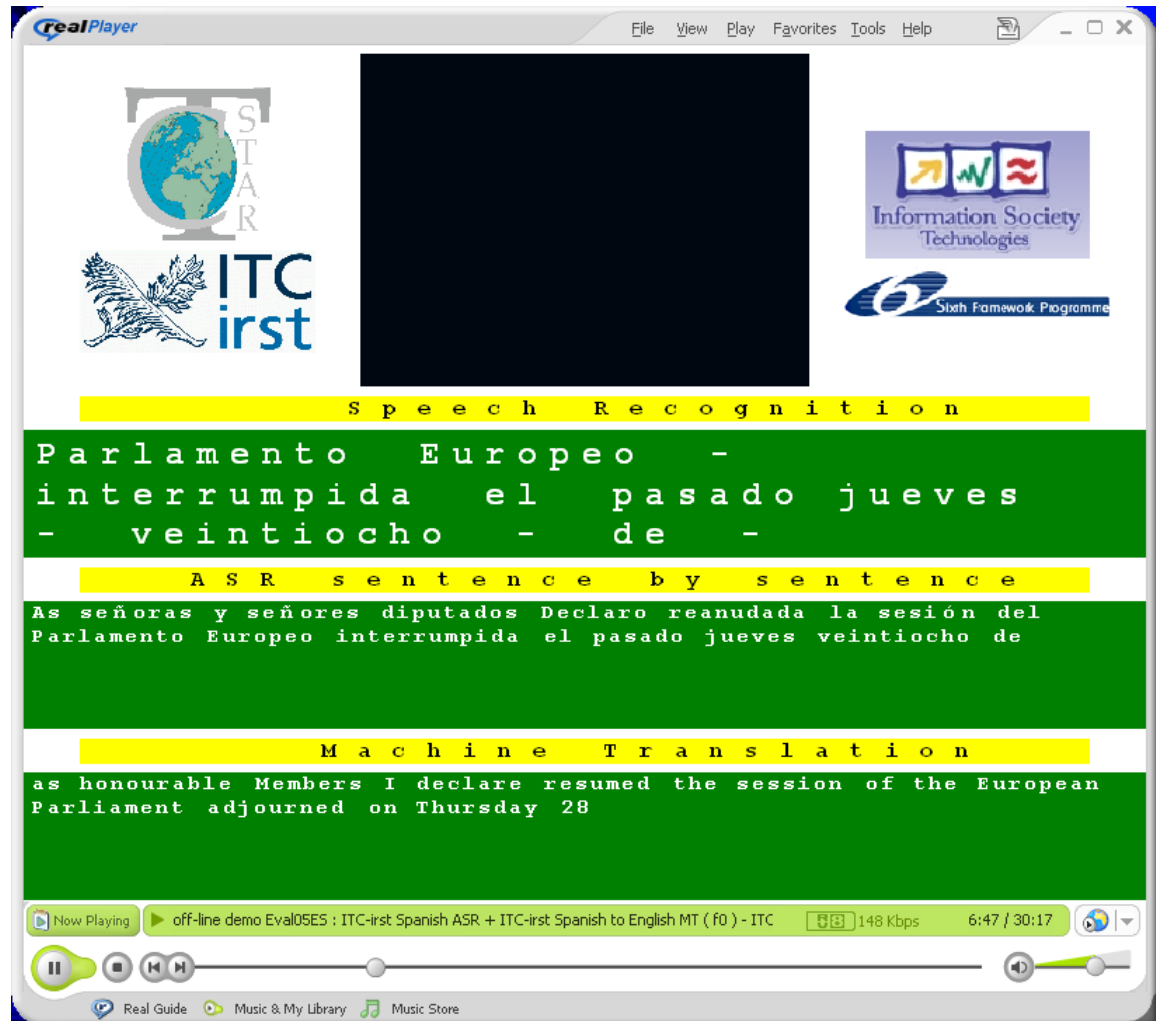
## ***Ejemplo experimental #2: PROYECTO TC-STAR***

- <http://www.tc-star.org>
- Integración de sistemas de reconocimiento de voz (ASR), traducción automática (MT) y síntesis de voz (TTS)
- Datos reales y de dominio amplio: transcripciones oficiales de las Sesiones Plenarias del Parlamento Europeo (EPPS)

	<b>Oraciones</b>	<b>Palabras</b>	<b>Vocabulario</b>
<b>Inglés</b>	<b>1.220.000</b>	<b>33.400.000</b>	<b>105.000</b>
<b>Castellano</b>	<b>1.220.000</b>	<b>34.800.000</b>	<b>169.000</b>



## Vídeo de demostración



The screenshot shows a RealPlayer window with a video player interface. The video content displays several logos at the top: TALP, ITC-irst, Information Society Technologies, and Sixth Framework Programme. The main video area is black, with text overlays on a green background. The text is as follows:

**S p e e c h   R e c o g n i t i o n**

Parlamento Europeo -  
interrumpida el pasado jueves  
- veintiocho - de -

**A S R   s e n t e n c e   b y   s e n t e n c e**

As señoras y señores diputados Declaro reanudada la sesión del  
Parlamento Europeo interrumpida el pasado jueves veintiocho de

**M a c h i n e   T r a n s l a t i o n**

as honourable Members I declare resumed the session of the European  
Parliament adjourned on Thursday 28

At the bottom, the player shows a 'Now Playing' bar with the title 'off-line demo Eval05ES : ITC-irst Spanish ASR + ITC-irst Spanish to English MT ( f0 ) - ITC', a bitrate of 148 Kbps, and a duration of 6:47 / 30:17. Playback controls (play, stop, previous, next, volume) and a progress slider are visible.

## Ejemplo experimental #3: PROYECTO TC-STAR

- Demostrador en línea para la traducción estadística entre castellano y catalán (datos reales y de dominio amplio)



### N-II: a statistical machine translator between Spanish and Catalan

This machine translation system is based on an N-gram translation model integrated in an optimized log-linear combination of additional features. The demo provides translation between the following pairs of languages:

Text to translate (up to 100 words)

Esta es una prueba del traductor automático entre castellano y catalán desarrollado por la UPC.

Spanish -> Catalan    Catalan -> Spanish    Clear

Translation:

Aquesta és una prova del traductor automàtic entre castellà i català desenvolupat per la UPC.

<http://www.n-ii.org>

## ***Ejemplo experimental #4: PROYECTO CHI-SPA\_MTAC***

- <http://www.talp.upc.edu/talp/>
- Sistema de traducción estadística entre chino y castellano
- Herramientas para la comunicación bilingüe entre chino y castellano:
  - 1.- traducción asistida
  - 2.- navegación bilingüe en Internet
  - 3.- mensajería electrónica (e-mail, SMS, chat)
  - 4.- video-conferencia bilingüe

## Resultados preliminares

请给我看看菜单。 **Le ruego me demuestran el menú , por favor .**  
(Please show me a menu)

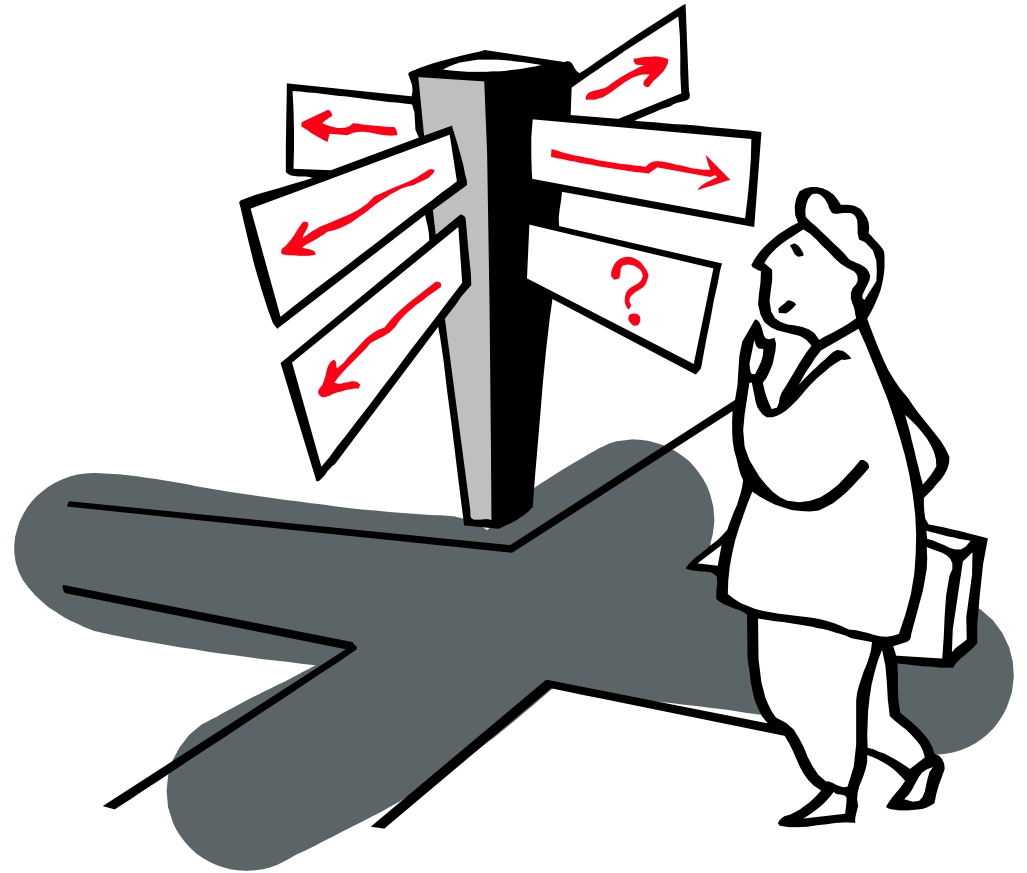
我想要导游。 **Me gustaría recibir una guía , por favor .**  
(I want to have a travel guide)

请给我一杯白葡萄酒。 **Le ruego me conceda un vaso blanco vino .**  
(Please give me a glass of white wine)

请稍等。 **Hace un momento , por favor .**  
(Just a moment please)

请叫服务生搬行李。 **Por favor , pedimos bellboy llevar el equipaje .**  
(Call someone to carry my bags please)

# *Retos futuros*

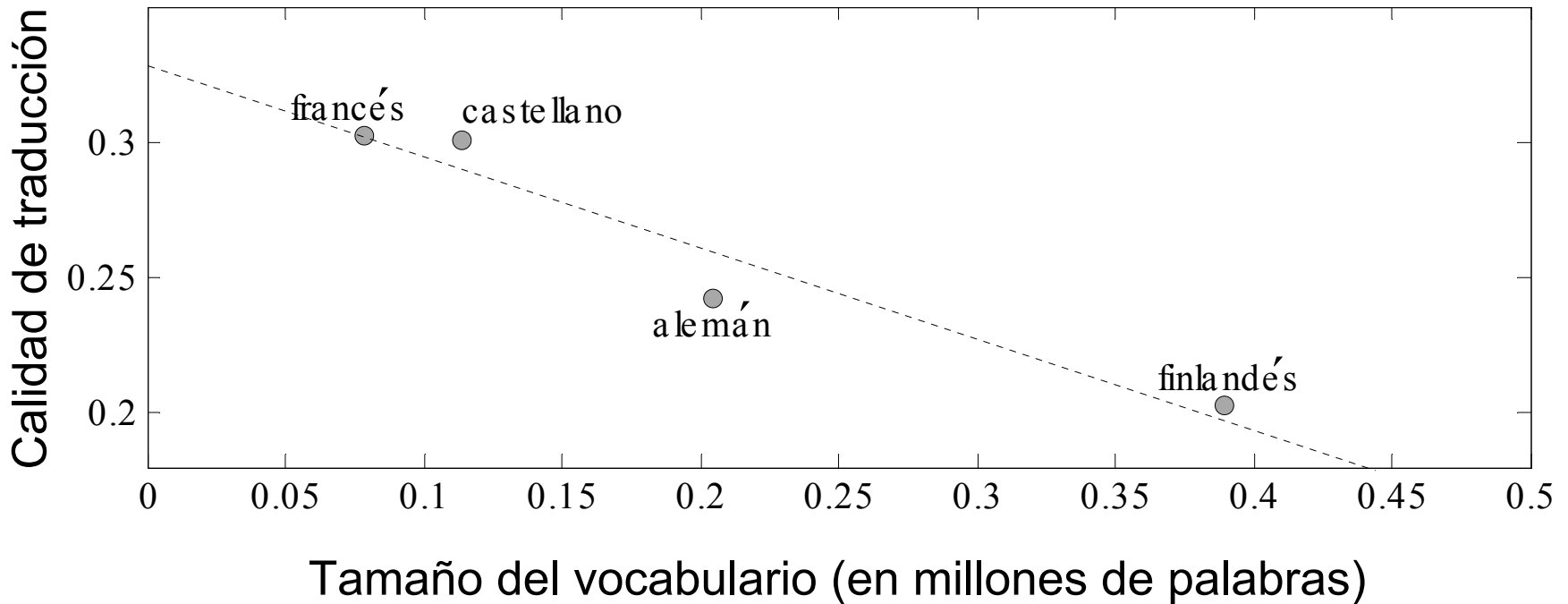


## ***Problemas específicos de la traducción automática***

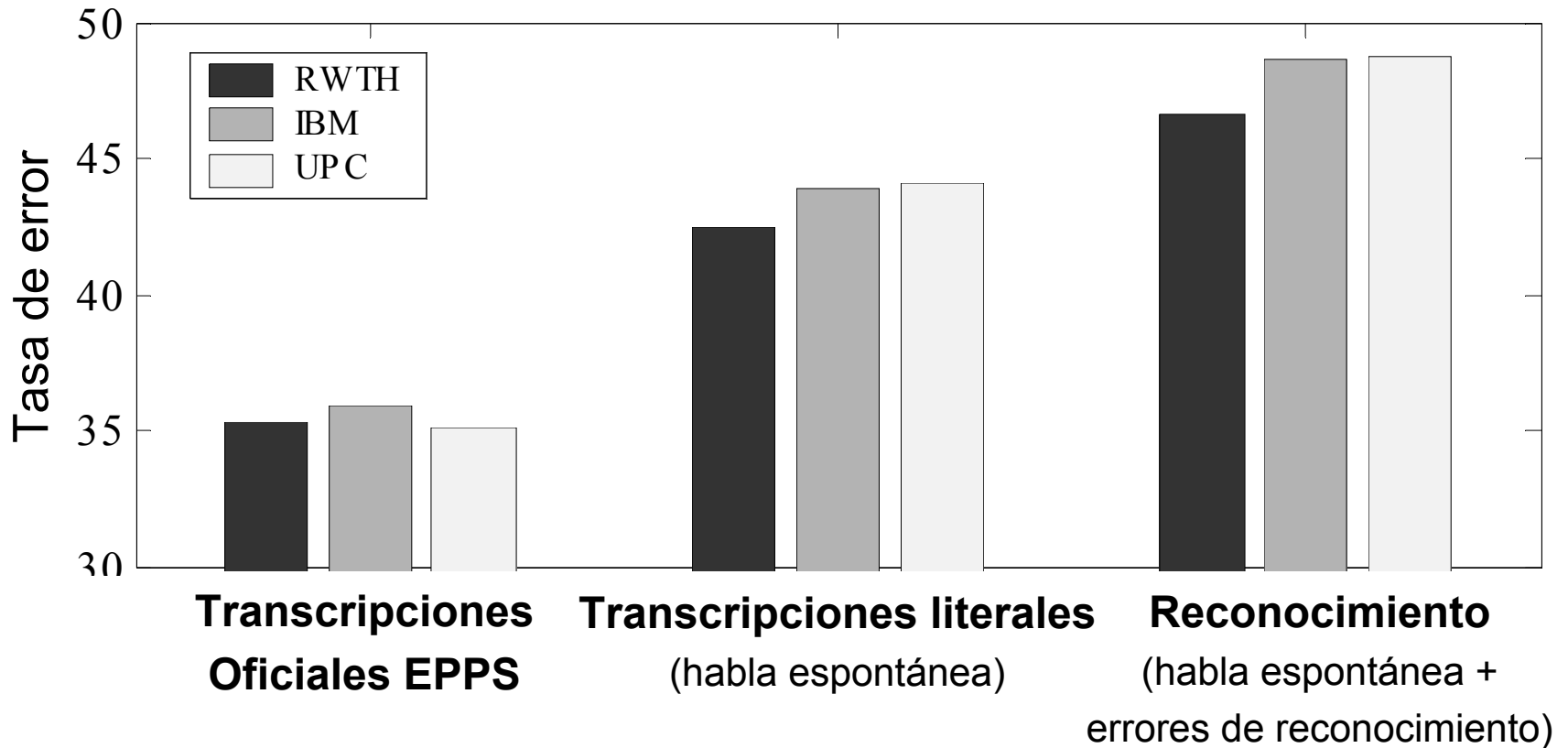
- 1.- Morfología: incide directamente sobre el tamaño del vocabulario, lo cual genera problemas de dispersión de los datos.
- 2.- Ordenamiento: afecta las traducciones entre lenguas gramaticalmente distantes, es un problema muy costoso desde el punto de vista computacional.
- 3.- Lenguaje oral: afecta la calidad de traducción debido a los distintos efectos del habla espontánea.



## ***Comparación de la calidad de traducciones al inglés desde cuatro lenguas fuente diferentes***



## Efectos del lenguaje oral en traducción





**Muchas Gracias  
por vuestra atención...**



Innovación y Conocimiento en la Sociedad Digital

# EDICIÓN 8ª Internet Global Congress

Barcelona, 29 de mayo - 1 de junio, 2006

PALACIO DE CONGRESOS, FIRA BARCELONA, PZA. DE ESPAÑA

La Investigación y el  
 Desarrollo en Tecnologías de  
 Traducción Automática

Rafael E. Banchs, UPC

Conferencia



[www.igcweb.net](http://www.igcweb.net)

UN PROYECTO DE:



Fundació  
 Barcelona  
 Digital