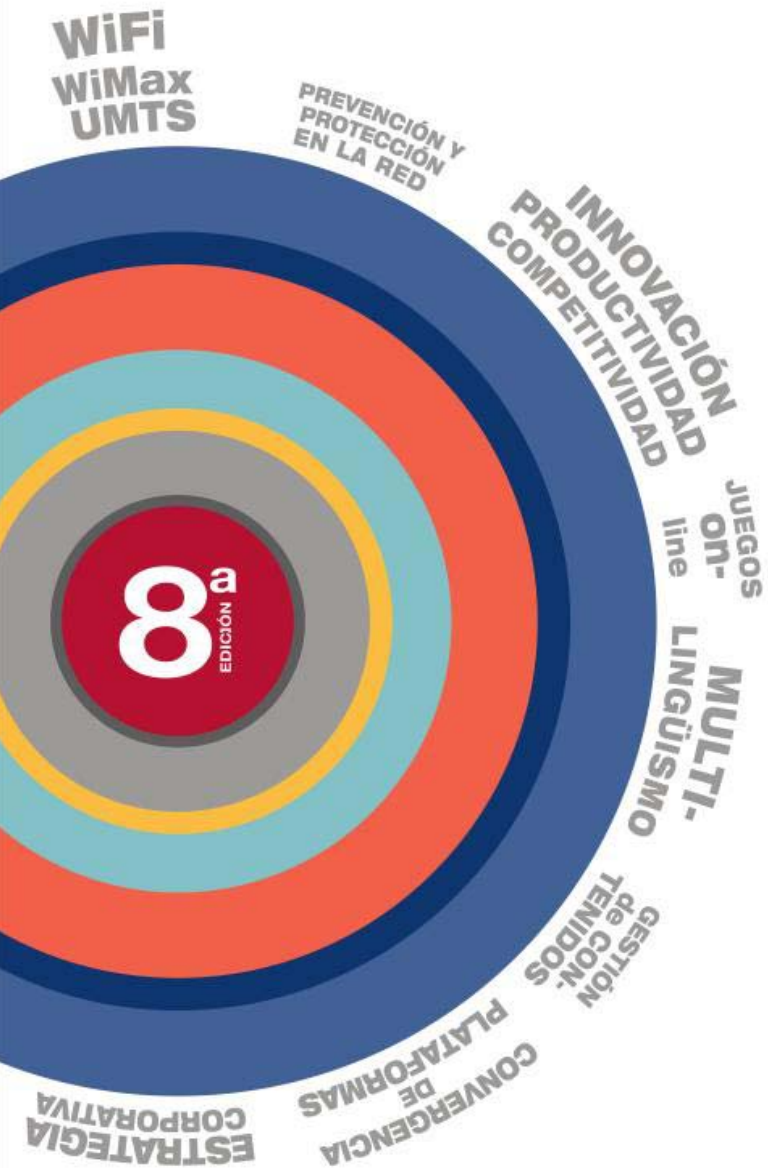


Innovación y Conocimiento en la Sociedad Digital

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

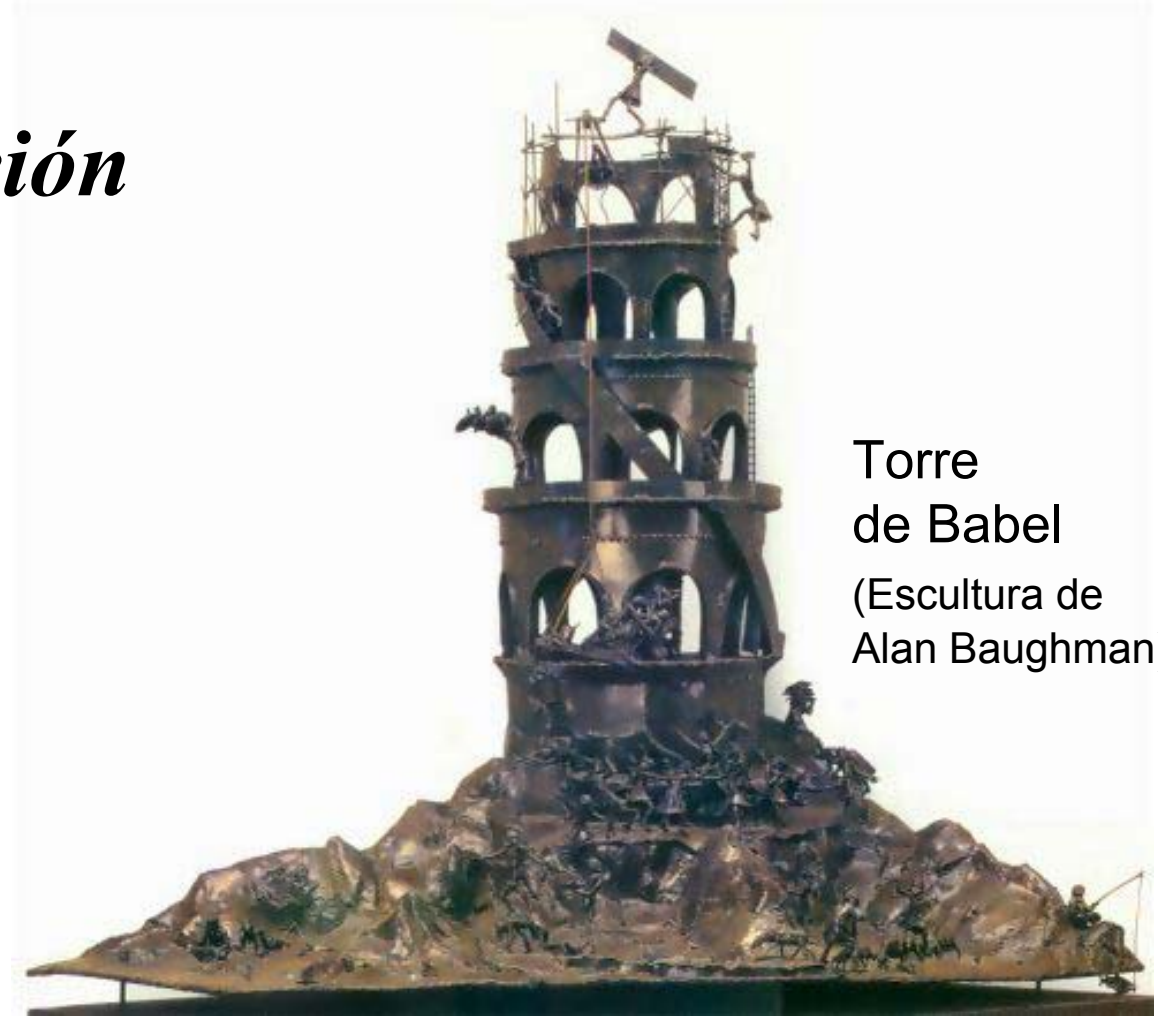
IGC

www.igcweb.net

UN PROYECTO DE:

Fundació Barcelona Digital

Introducción



Torre
de Babel
(Escultura de
Alan Baughman)

La traducción automática como área de investigación

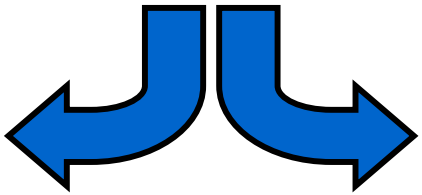
Búsqueda en *www.google.es*	*Resultados*
“ machine translation ”	***559.000***
“ machine translation ” + research	***196.000***
“ machine translation ” + research + university	***131.000***
“ machine translation ” + research – university	***63.800***
“ machine translation ” + confenerce	***123.000***
“ machine translation ” + journal	***98.100***

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

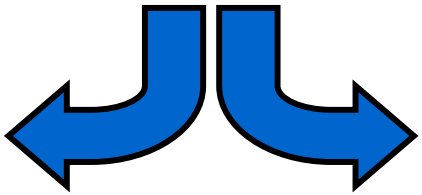
- Interlingua
  - Transfer
- Traducción directa

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

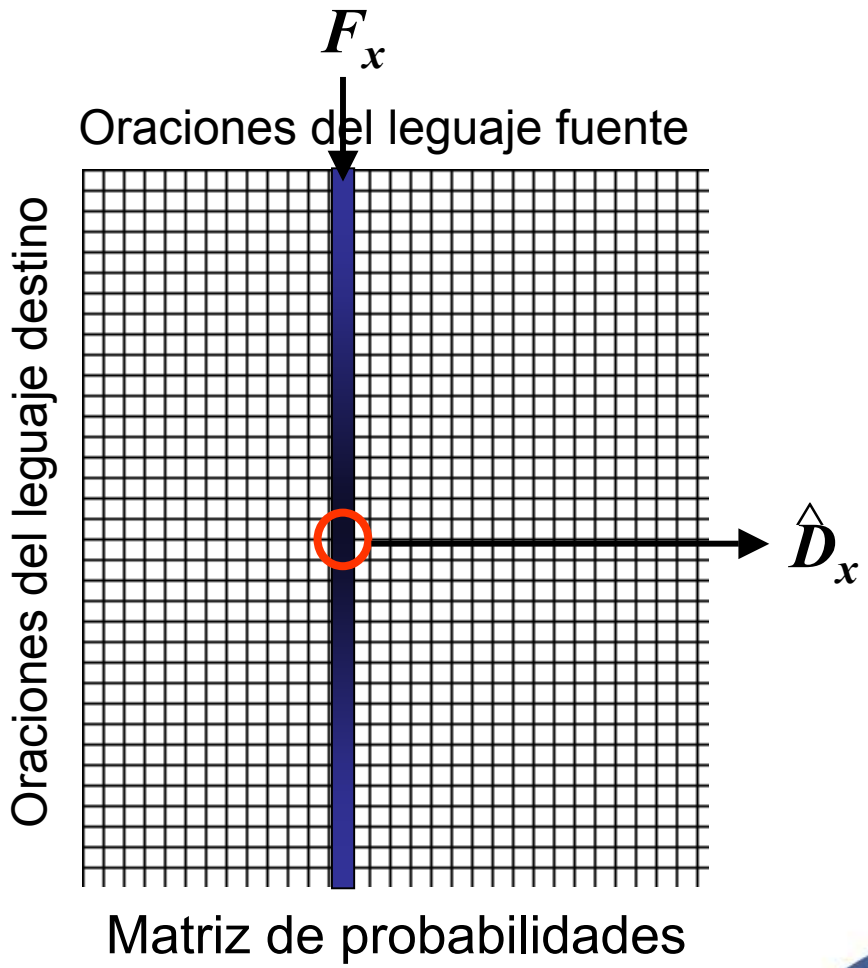
La aproximación estadística

Planteamiento teórico de la aproximación estadística

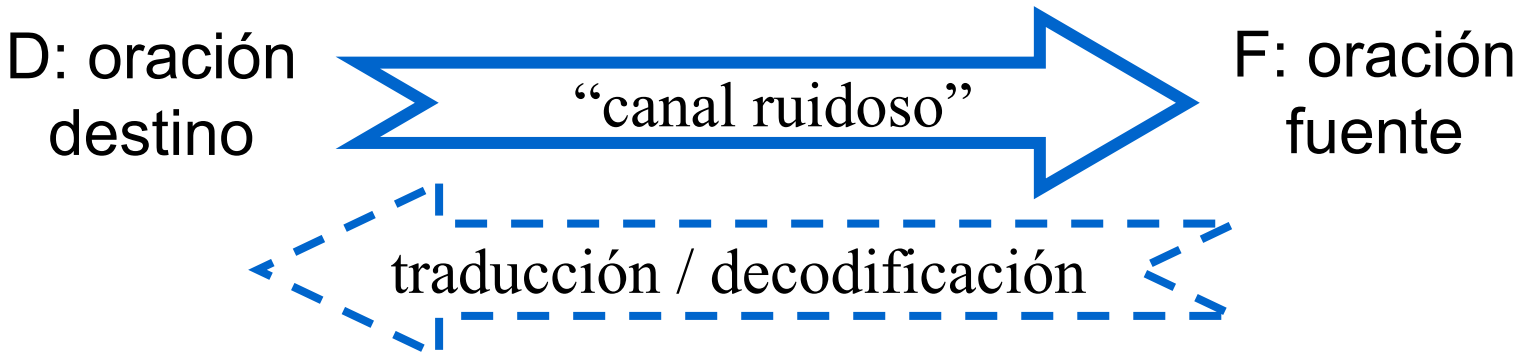
$$\hat{D}_x = \underset{D}{argmax}\ P(D|F_x)$$

Problemas prácticos:

- *Cálculo de las probabilidades*
- *Espacio de búsqueda*



Primer modelo de traducción estadística



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

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

Oración	**Probabilidad***
“the welcome mr. to sesion president”	***4,52x10^{-8} (-16,91)***
“mr. president welcome to the sesion”	***1,11x10^{-6} (-13,71)***
“sesion the to president mr. welcome”	***5,02x10^{-9} (-19,11)***
“president the sesion to welcome mr.”	***6,23x10^{-8} (-16,59)***
“sesion president welcome to mr. the”	***8,96x10^{-9} (-18,53)***

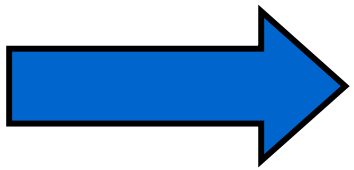
** Probabilidades calculadas con un modelo de 3-gramas entrenado con datos del Parlamento Europeo.*

Ejemplo con modelo de traducción basado en palabras

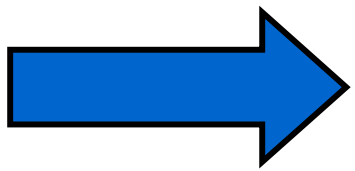
Oración fuente	**Probabilidad***
"el presidente vino al parlamento"	
Oraciones destino	
"the president wine to parliament"	$7{,}35 \times 10^{-7}$ (-14,12)
"the president came to parliament"	$3{,}42 \times 10^{-8}$ (-17,19)
"the parliament came to president"	$3{,}42 \times 10^{-8}$ (-17,19)

** 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)*

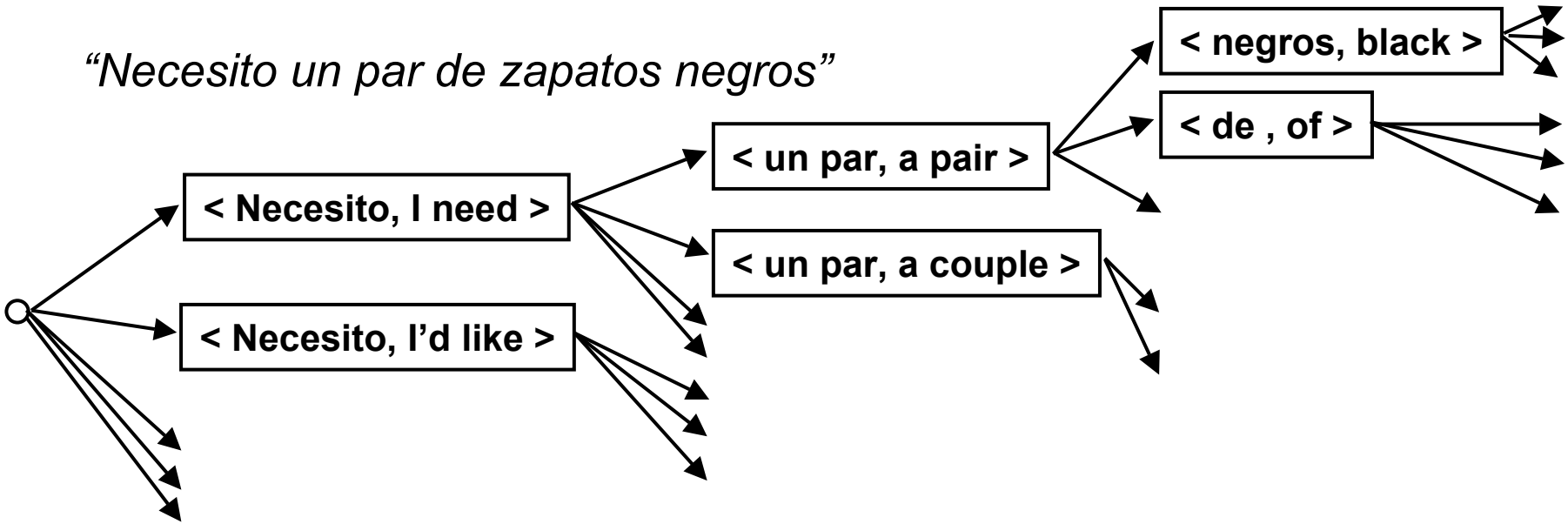
$$\hat{D} = \underset{D}{argmax}\ p(D|F) \approx \underset{D}{argmax} \prod_i p_i(F,D)^{\lambda_i}$$

Canal Ruidoso ⇨ caso particular:

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

Decodificación con modelos basados en "Frases"

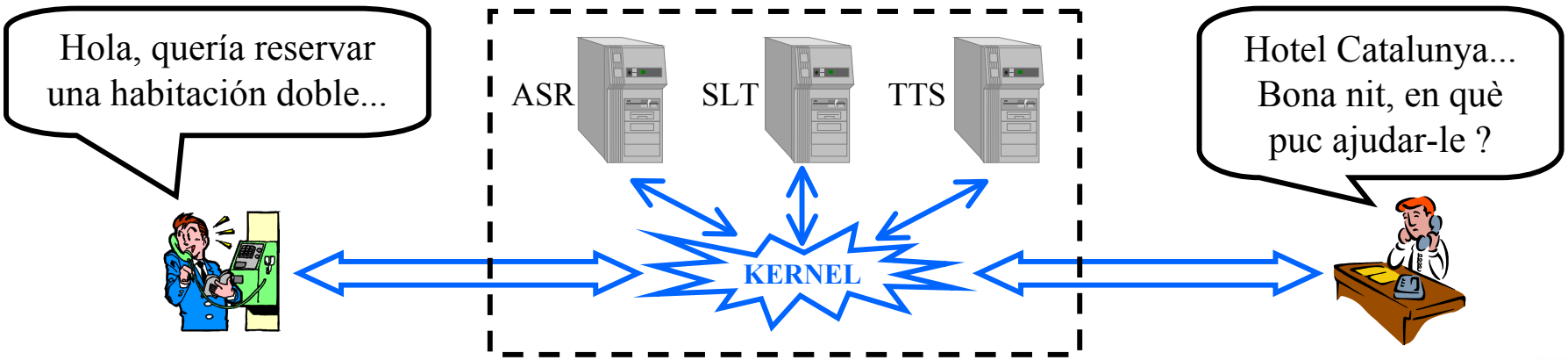
Se explora el espacio de las posibles traducciones mediante el uso de un algoritmo de búsqueda *(Wang y Waibel 1997, Tillman et al. 1997, Koehn 2004)*



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

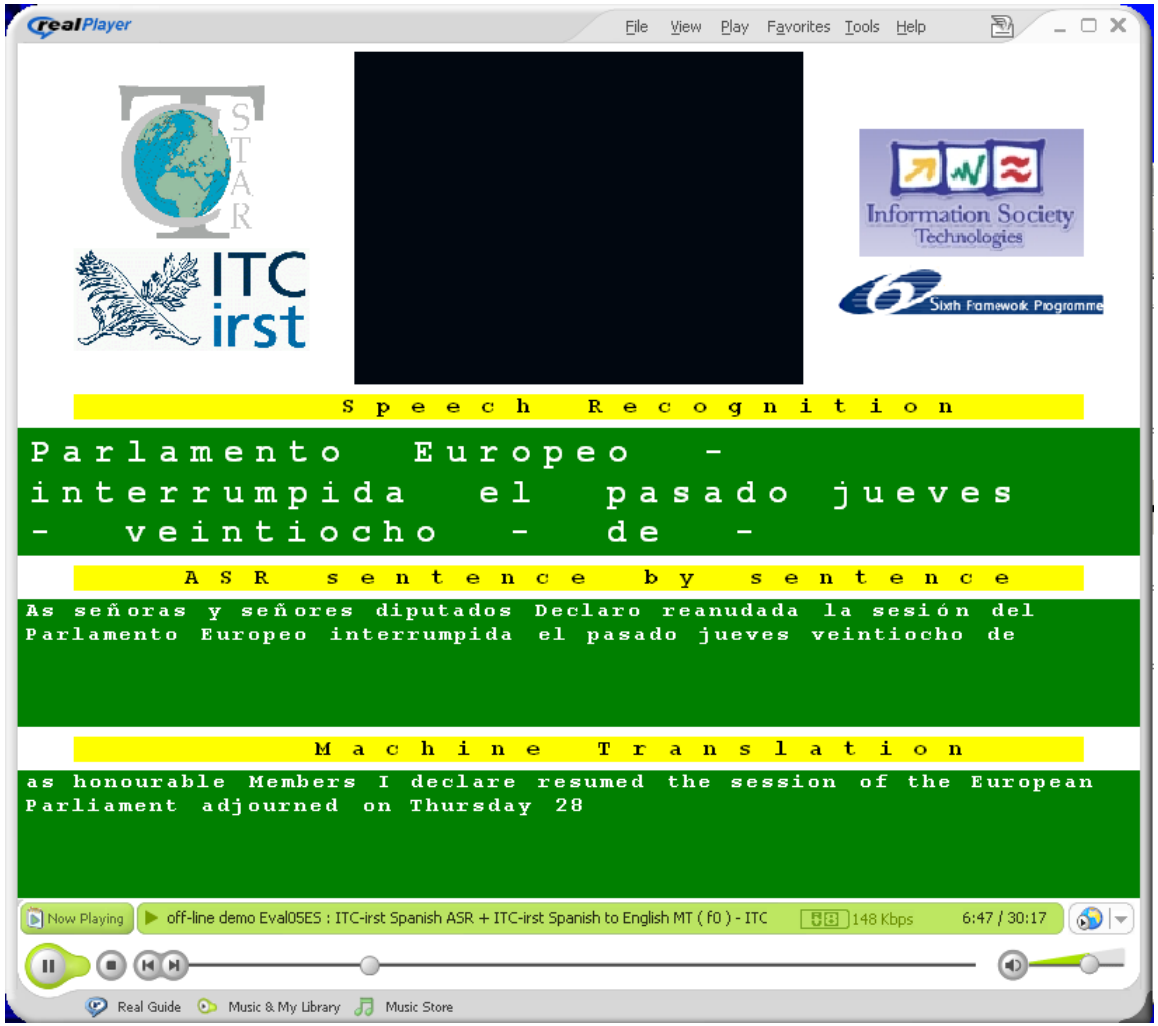
Objetivos	dlg1	dlg2	dlg3	dlg4	dlg5	dlg6	obj
día de arribo	0.75	0.60	1.00	-0.50	0.50	-0.63	**0.29**
noches	1.00	0.38	1.00	1.00	0.50	1.00	**0.81**
habitación	1.00	1.00	0.50	0.50	0.50	0.43	**0.65**
precio	-0.50	1.00	1.00	0.60	0.60	0.27	**0.50**
nombre	–	0.60	1.00	1.00	0.50	0.60	**0.74**
tipo tc	1.00	1.00	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**
diálogo	**0.61**	**0.45**	**0.70**	**0.47**	**0.47**	**0.03**	

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)

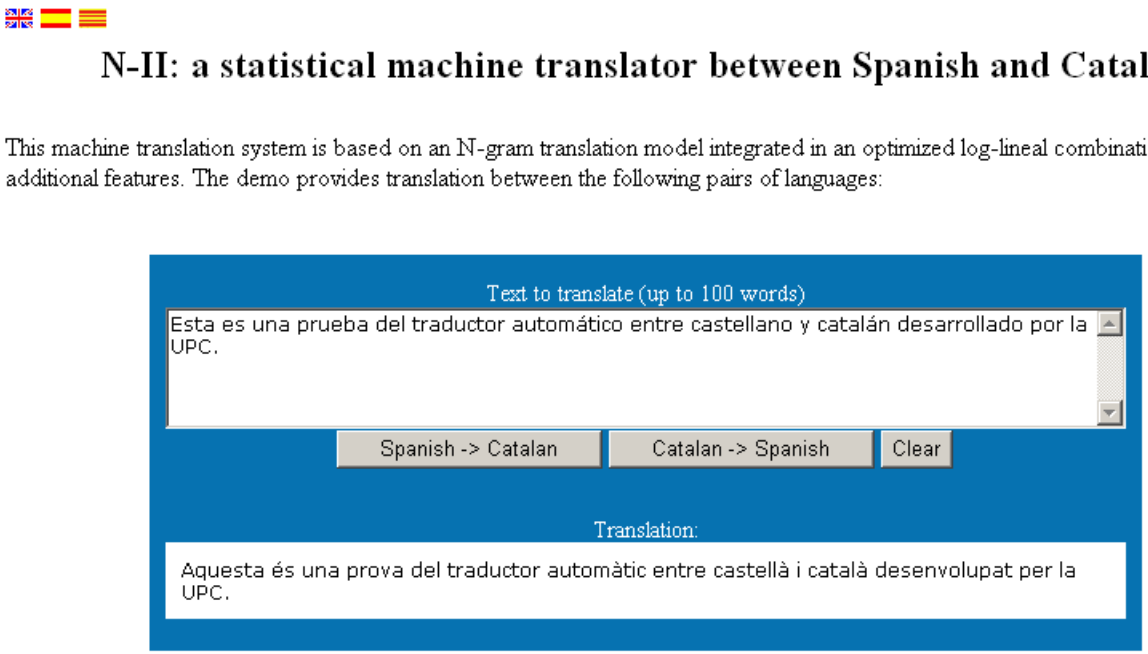
	Oraciones	Palabras	Vocabulario
Inglés	**1.220.000**	**33.400.000**	**105.000**
Castellano	**1.220.000**	**34.800.000**	**169.000**

Vídeo de demostración



Ejemplo experimental #3: PROYECTO TC-STAR

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



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)

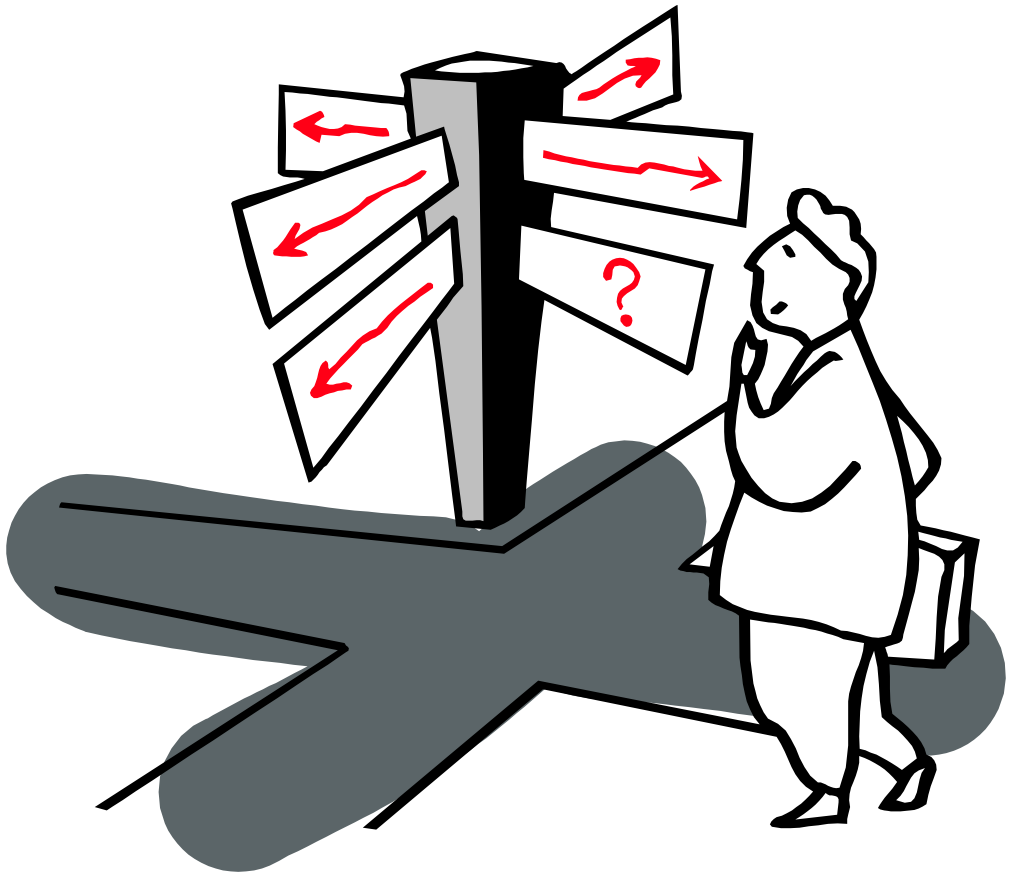
我 想要 导游 。 **Me gustaría recibir una guía , por favor .**
(I want to have a travel guide)

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



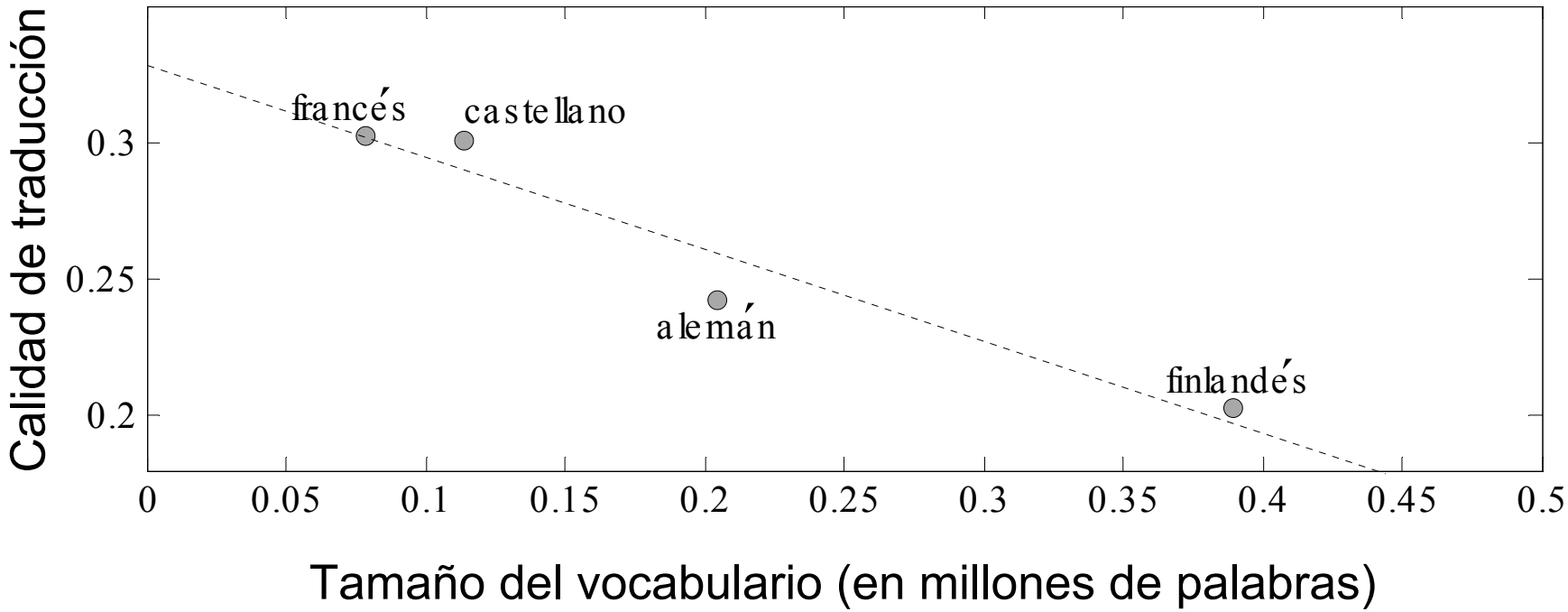
Problemas específicos de la traducción automática

1.- Morfología: incide directamente sobre el tamaño del vocabulario, lo cual genera problemas de dispersión de los datos.

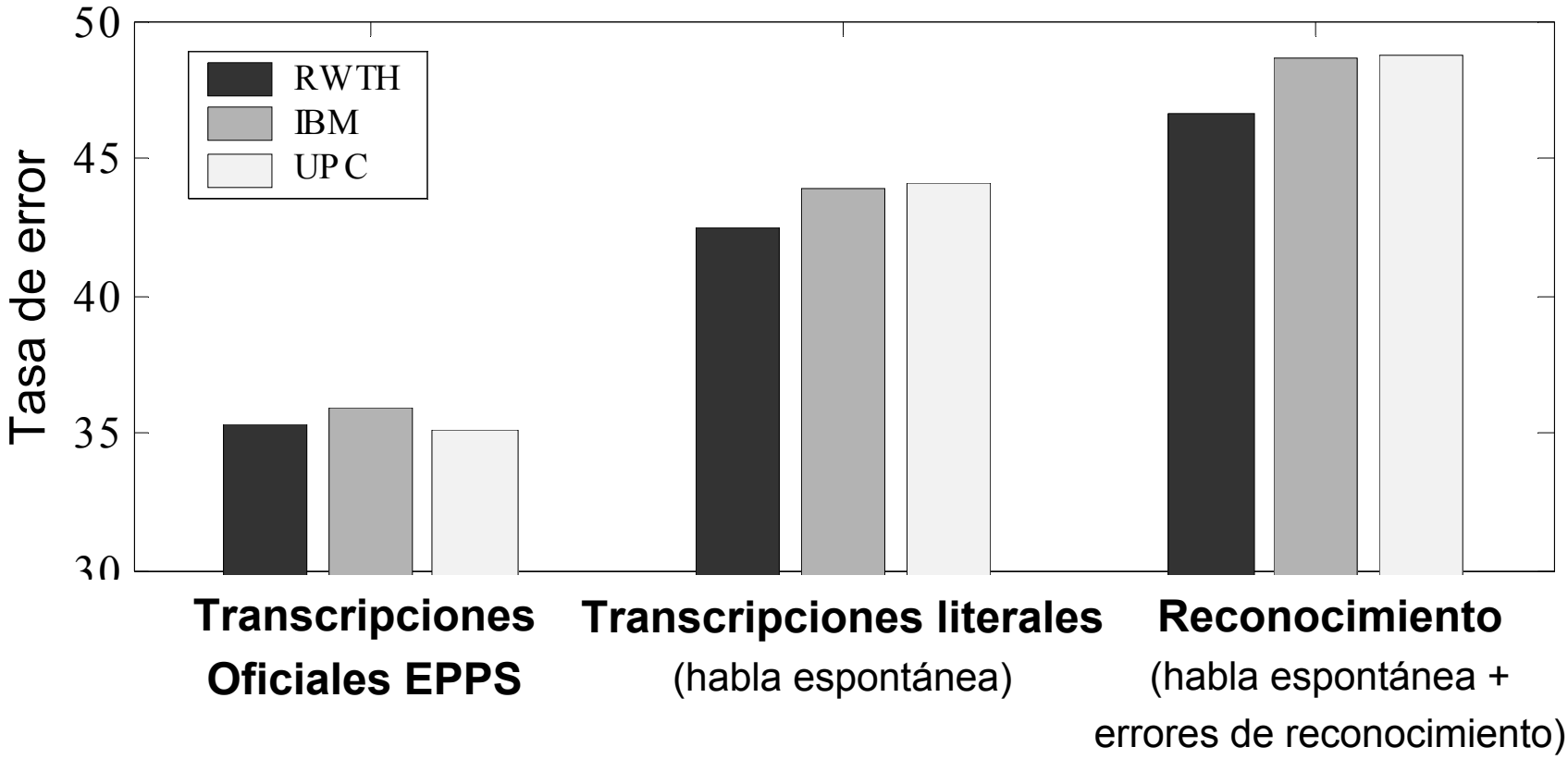
2.- Ordenamiento: afecta las traducciones entre lenguas gramaticalmente distantes, es un problema muy costoso desde el punto de vista computacional.

3.- Lenguaje oral: afecta la calidad de traducción debido a los distintos efectos del habla espontánea.

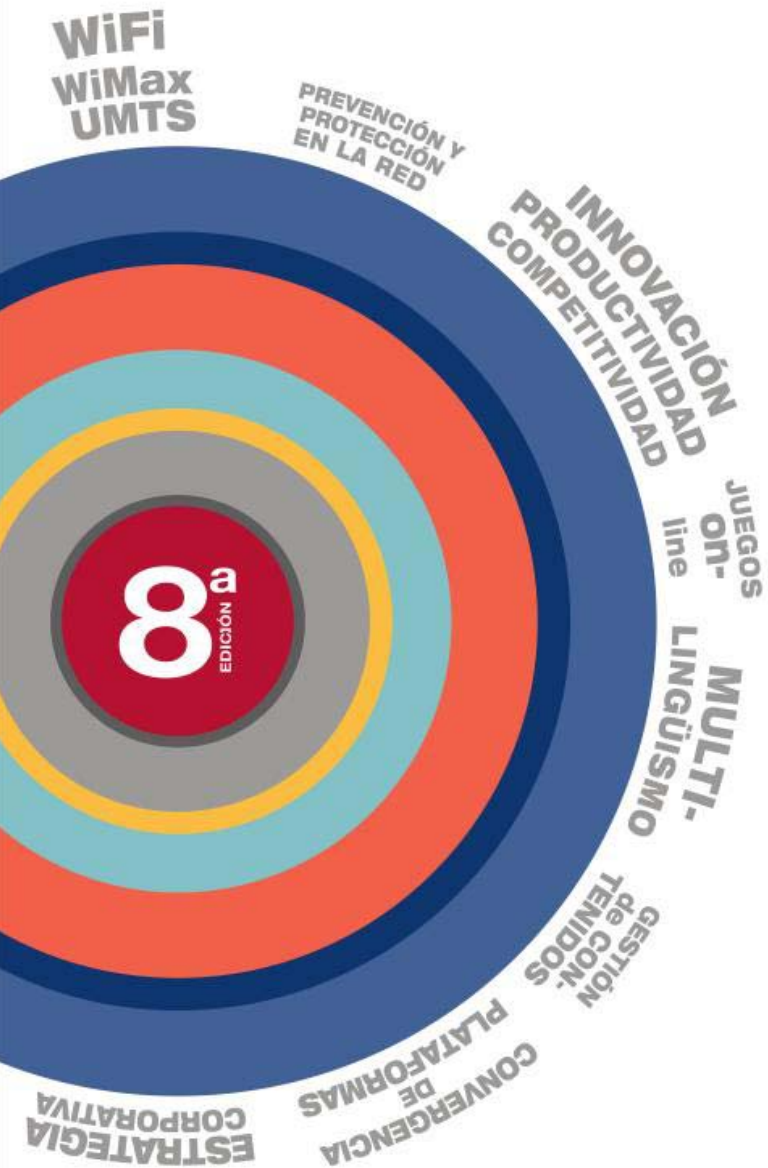
Comparación de la calidad de traducciones al inglés desde cuatro lenguas fuente diferentes



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

IGC

www.igcweb.net

UN PROYECTO DE:

Fundació Barcelona Digital