UNIVERSIDAD CATOLICA ANDRÉS BELLO
Facultad de Ingeniería
Escuela de Telecomunicaciones

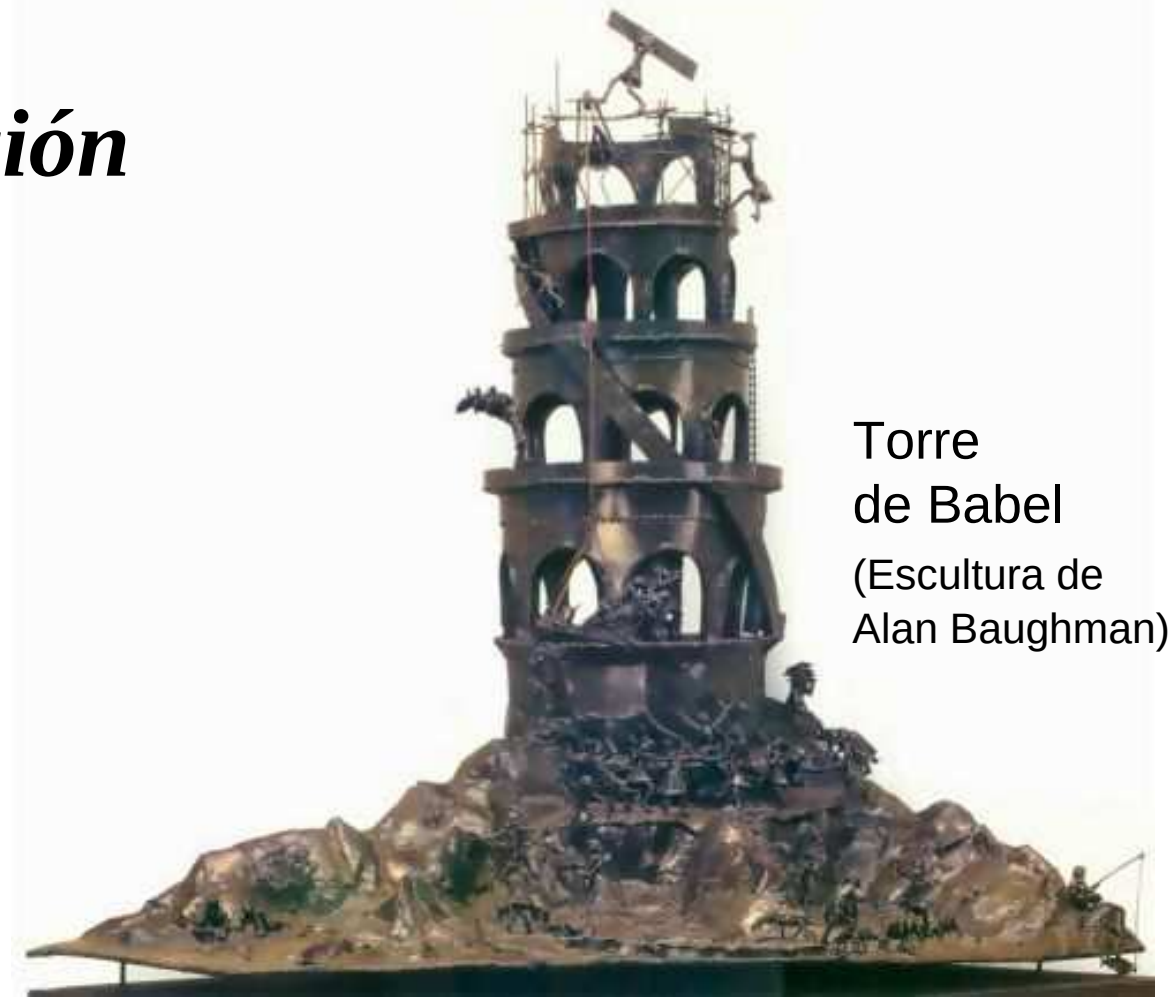
Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Traducción Automática Estadística

Rafael E. Banchs
Unversitat Politècnica de Catalunya

V Jornadas de Ingeniería de Telecomunicaciones – Universidad Católica Andrés Bello – Caracas, 10 de Mayo de 2007

Introducción



Torre
de Babel
(Escultura de
Alan Baughman)

Evolución de la traducción automática

- Siglo IV: San Jerónimo traduce la Biblia al latín.
- Siglo XVII: Primeros intentos de desarrollo de lenguajes universales.

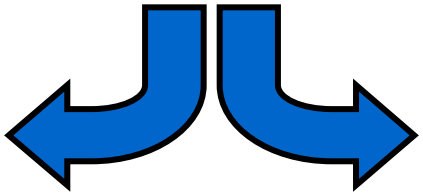
- Siglo XX: Aparición del ordenador
    - 40s: La traducción como problema criptográfico.
    - 70s: Un trabajo para la inteligencia artificial.
    - 90s: Nacimiento de la traducción automática estadística.
- Siglo XXI: Aplicaciones prácticas de los sistemas de traducción ???

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***

Dos paradigmas, cinco métodos

Métodos de Traducción Automática



Basados en Conocimiento

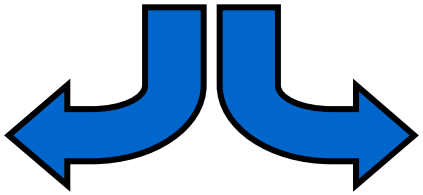
- Interlingua
- Transfer
- Traducción directa

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

UNIVERSIDAD CATOLICA ANDRÉS BELLO
Facultad de Ingeniería
Escuela de Telecomunicaciones

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
UNIVERSITAT POLITÈCNICA DE CATALUNYA

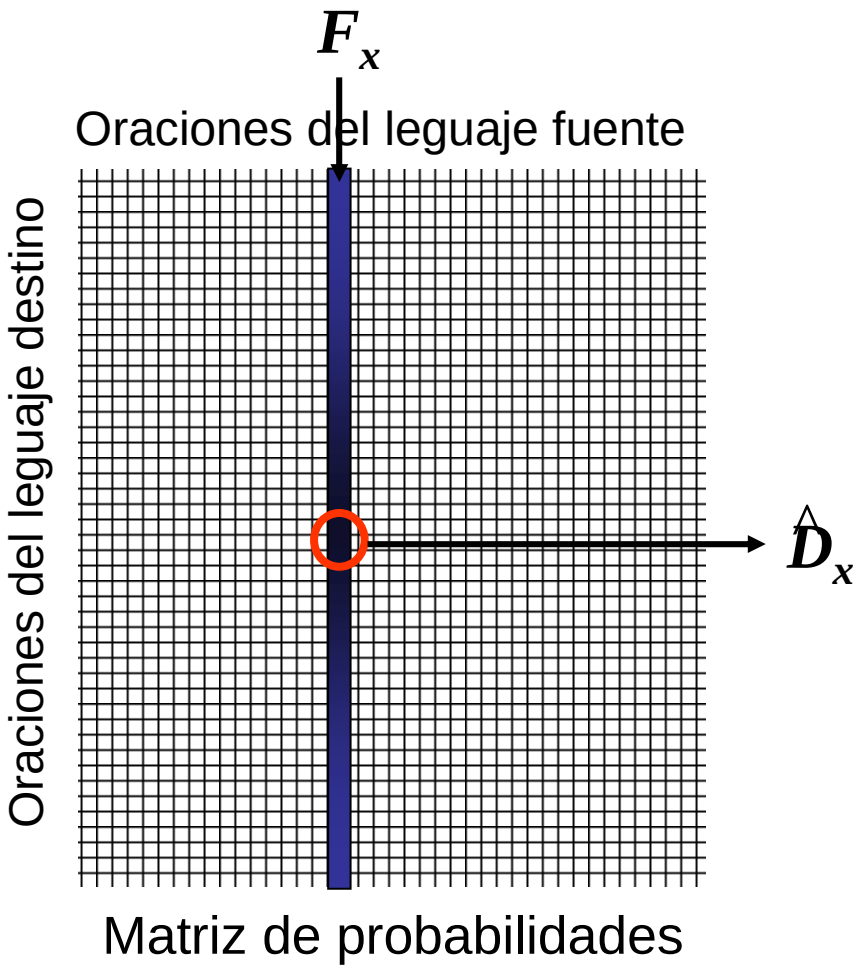
La aproximación estadística

Planteamiento teórico de la aproximación estadística

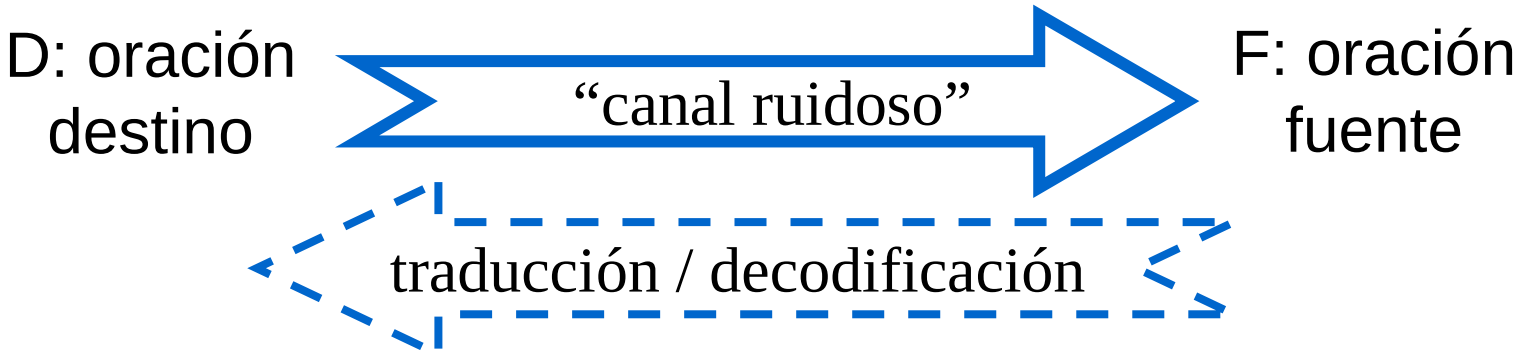
$$\hat{D}_x = \underset{D}{argmax}\ P(D|F_x)$$

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 $\boldsymbol{P(D|F)}$ se convierte en la optimización simultánea de dos características:

1.- *"Adequacy"*: búsqueda de los contenidos más adecuados de acuerdo con el modelo de traducción $\boldsymbol{P(F|D)}$

2.- *"Fluency"*: búsqueda de la mejor construcción gramatical de acuerdo con el modelo de lenguaje $\boldsymbol{P(D)}$

Modelo del lenguaje destino: P(D)

Dada una oración $D: d_1\ d_2\ d_3 \ldots d_k$

La probabilidad de dicha oración está dada por:

$$p(D) = p(d_1, d_2, d_3 \ldots d_k)$$

$$p(D) = p(d_1)\ p(d_2 \mid d_1)\ p(d_3 \mid d_2, d_1) \ldots p(d_k \mid d_{k-1} \ldots d_2, d_1)$$

Aproximación del n-grama (generalmente n=3):

$$p(d_j \mid d_{j-1} \ldots d_2, d_1) \approx p(d_j \mid d_{j-1} \ldots d_{j-n+1})$$

Entrenamiento de un modelo de n-gramas

Los n-gramas son fáciles de entrenar a partir de un corpus de datos. Así por ejemplo para los 3-gramas:

$$p(d_j \mid d_{j-1}, d_{j-2}) \approx \frac{\textit{Número de veces: } d_j, d_{j-1}, d_{j-2}}{\textit{Número de veces: } d_{j-1} \ldots d_{j-2}}$$

Y la probabilidad de la oración se aproxima como:

$$p(D) = p(d_3 \mid d_2, d_1)\, p(d_4 \mid d_3, d_2)\, p(d_5 \mid d_4, d_3) \ldots p(d_k \mid d_{k-1}, d_{k-2})$$

Ejemplo de probabilidades usando 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.*

Modelo de traducción: $P(F|D)$

Los primeros modelos de traducción fueron los propuestos por *Brown et al. (1993)*:

1.- son modelos basados en palabras: $p(f_i|d_j)$

2.- un total de 5 modelos de complejidad creciente:

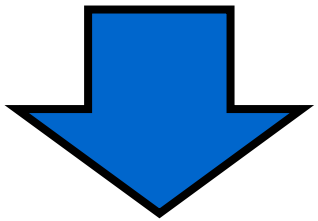
IBM1 ⇨ IBM2 ⇨ IBM3 ⇨ IBM4 ⇨ IBM5

3.- requieren la existencia de un corpus bilingüe alineado

Alineado de un corpus bilingüe

Necesito un par de zapatos negros

I need a pair of black shoes



Necesito un par de zapatos negros

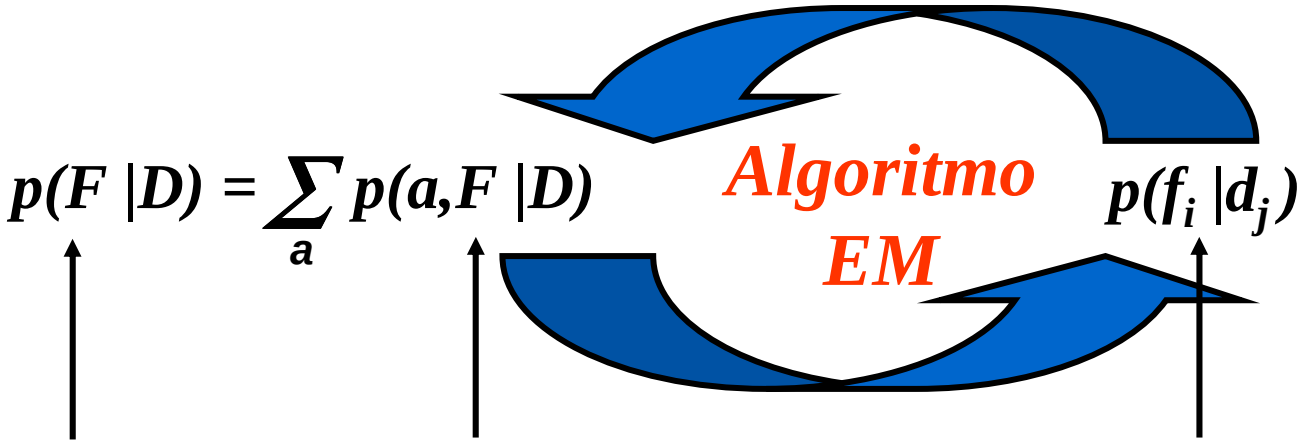
I need a pair of black shoes

Probabilidad de alineado

$$p(a,F\,|D)$$

Cálculo automático de parámetros

Los parámetros pueden ser calculados en forma recursiva usando el algoritmo EM desarrollado por *Baum (1972)*:



Ejemplo de probabilidades de traducción

Oración fuente	**Probabilidad***
"el presidente vino al parlamento"	
Oraciones destino	
"the president wine to parliament"	***7,35x10^{-7} (-14,12)***
"the president came to parliament"	***3,42x10^{-8} (-17,19)***
"the parliament came to president"	***3,42x10^{-8} (-17,19)***
"the engineer solved the problem"	***4,82x10^{-15} (-32,96)***

** 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:

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

Modelos de traducción basados en "Frases"

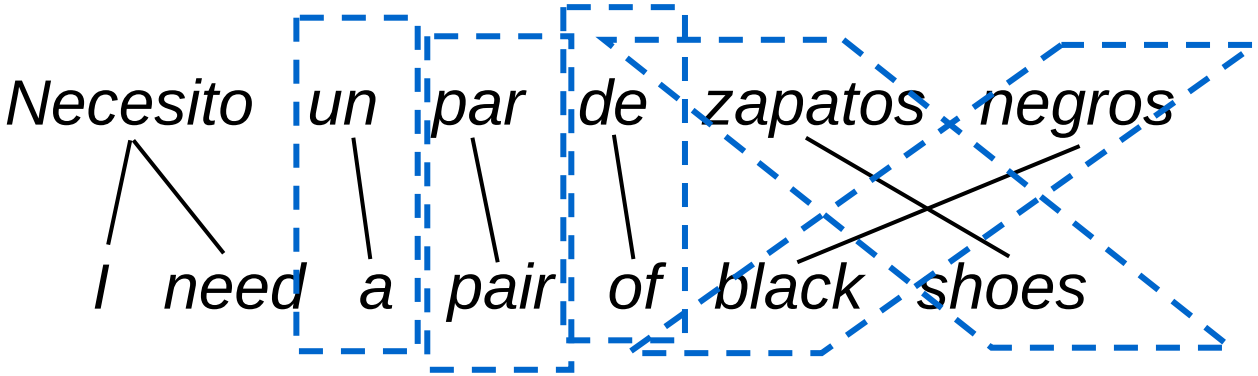
Frase: unidad bilingüe que contiene una o más palabras fuente y destino.

1.- todas la palabras dentro de un frase sólo pueden estar alineadas con palabras dentro de la misma frase,

2.- la longitud (número de palabras) suele restringirse a un máximo de 4 ó 5 palabras,

3.- las probabilidades se estiman a mediante conteo en el corpus de entrenamiento: $p(d|f) = N(d,f) / N(f)$

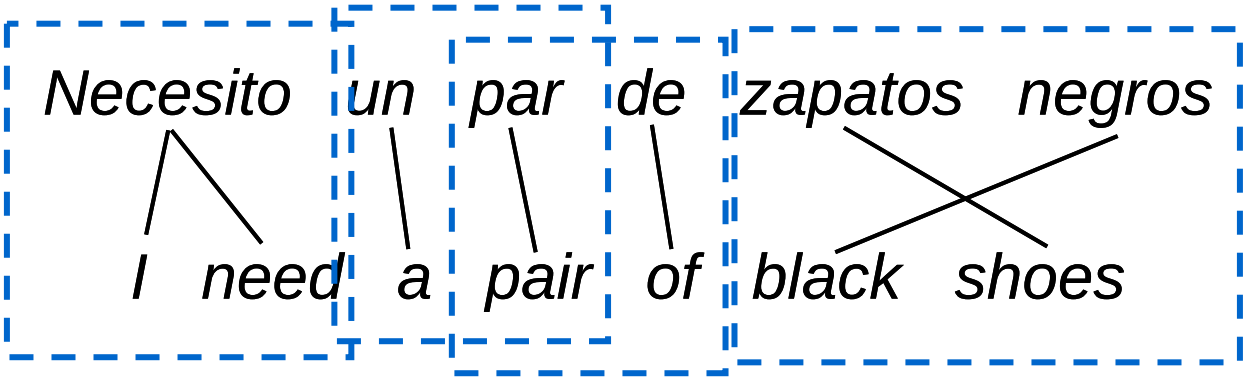
Ejemplo de extracción de "Frases"



Longitud 1

< un , a >
< par , pair >
< de , of >
< negros , black >
< zapatos , shoes >

Ejemplo de extracción de "Frases"



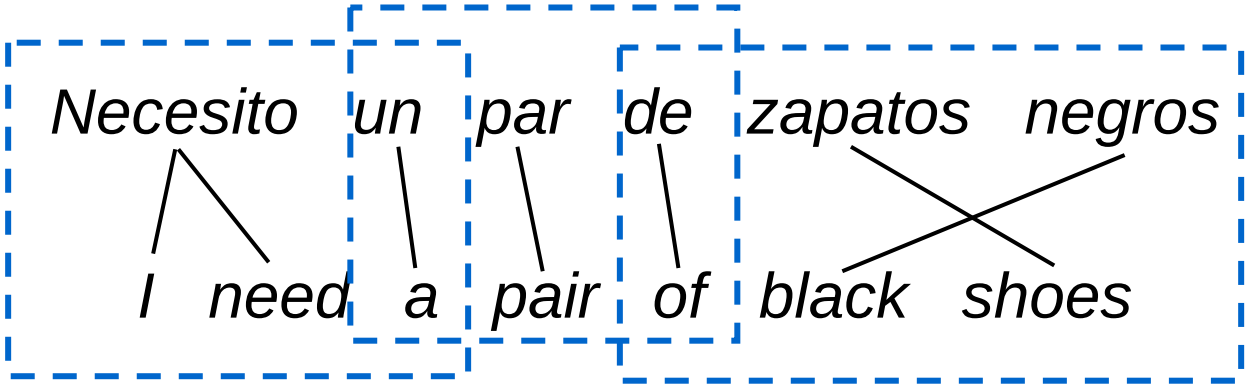
Longitud 1

< un , a >
< par , pair >
< de , of >
< negros , black >
< zapatos , shoes >

Longitud 2

< necesito , I need >
< un par , a pair >
< par de , pair of >
< zapatos negros , black shoes >

Ejemplo de extracción de "Frases"



Longitud 1
< un , a >
< par , pair >
< de , of >
< negros , black >
< zapatos , shoes >

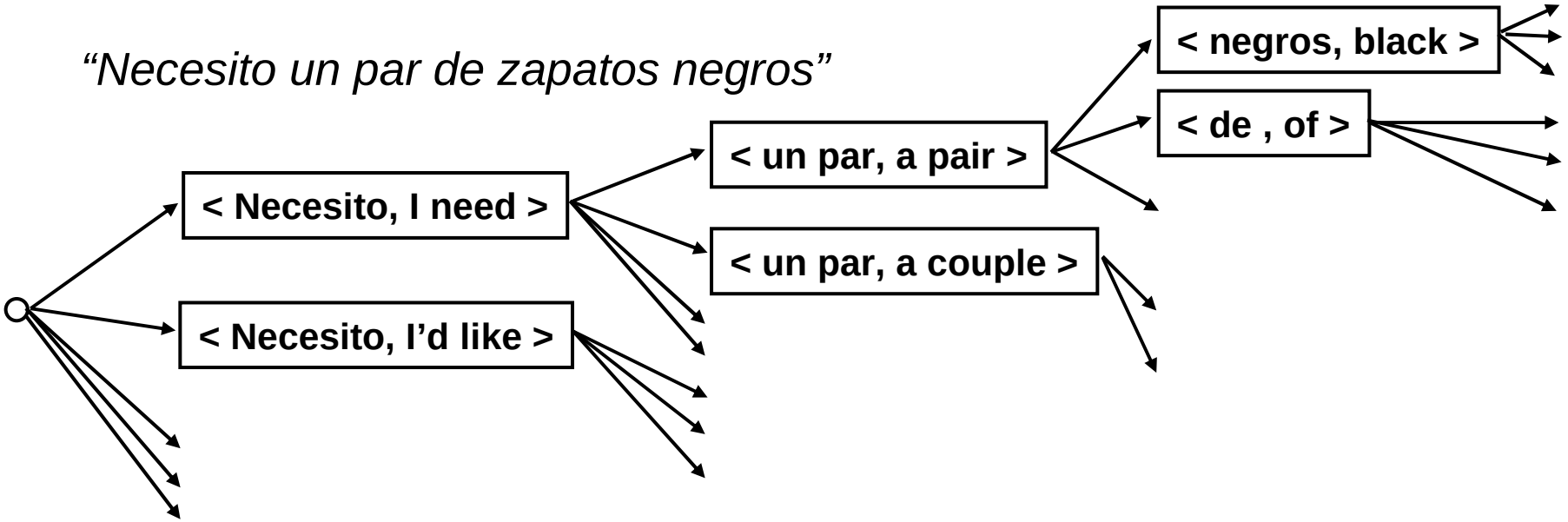
Longitud 3
< necesito un , I need a >
< de zapatos negros , of black shoes >
< un par de , a pair of >

Longitud 2
< necesito , I need >
< un par , a pair >
< par de , pair of >
< zapatos negros , black shoes >

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"



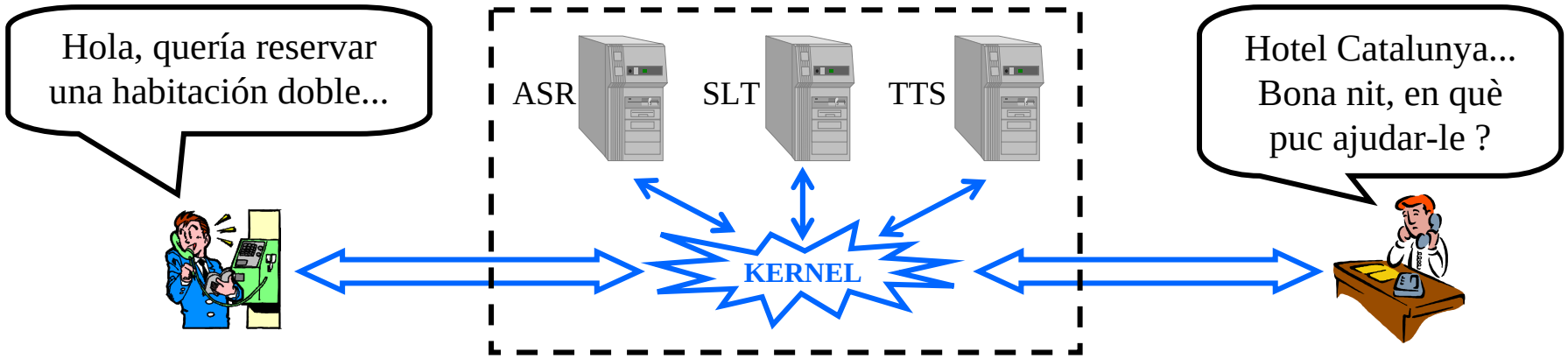
UNIVERSIDAD CATOLICA ANDRÉS BELLO
Facultad de Ingeniería
Escuela de Telecomunicaciones

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
UNIVERSITAT POLITÈCNICA DE CATALUNYA

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)

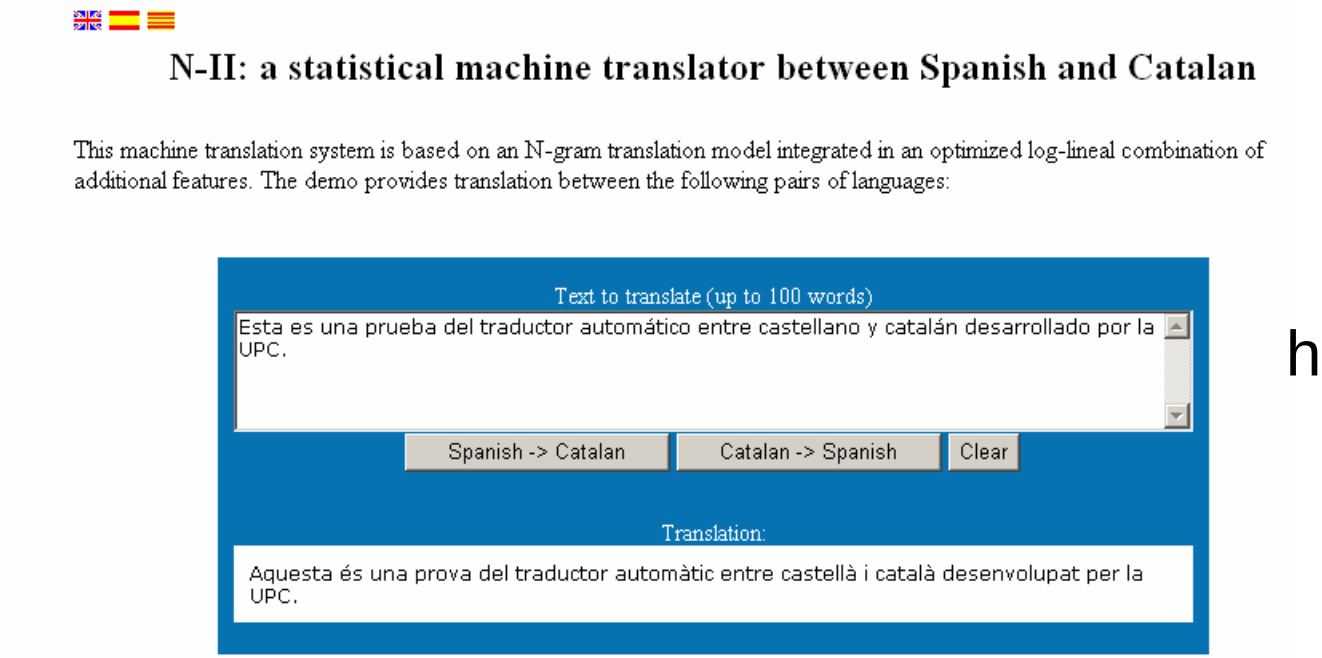
	Oraciones	**Palabras**	**Vocabulario**
Inglés	**1.220.000**	**33.400.000**	**105.000**
Castellano	**1.220.000**	**34.800.000**	**169.000**

Vídeo de demostración

- Procesado "*off-line*"
- Sistema de reconocimiento automático del habla (ASR) y traducción automática del habla (SLT).
- Dirección de traducción: Castellano ⇨ Inglés

Ejemplo experimental #3: PROYECTO TC-STAR

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

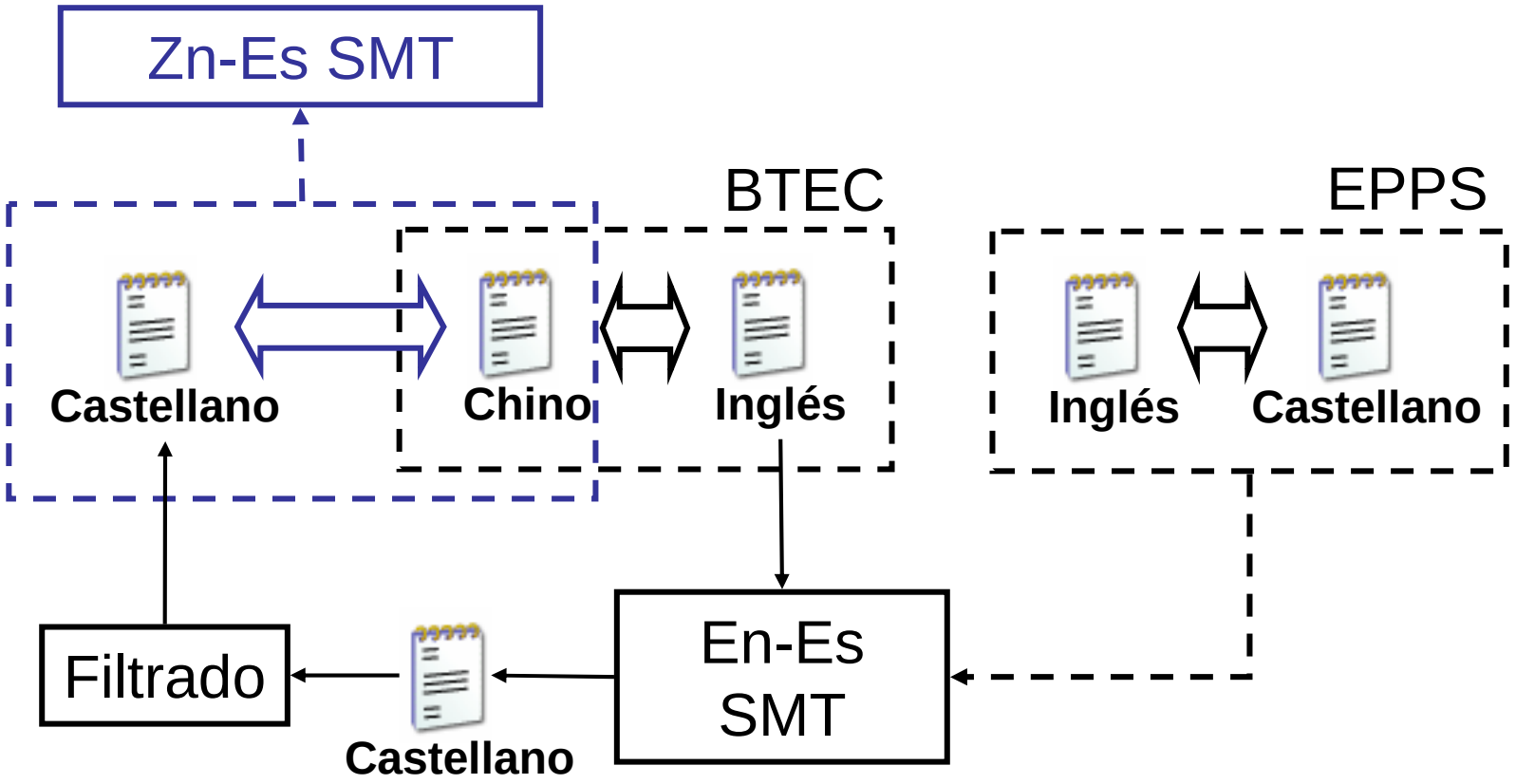


http://www.n-ii.org

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

Experimentación preliminar



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)

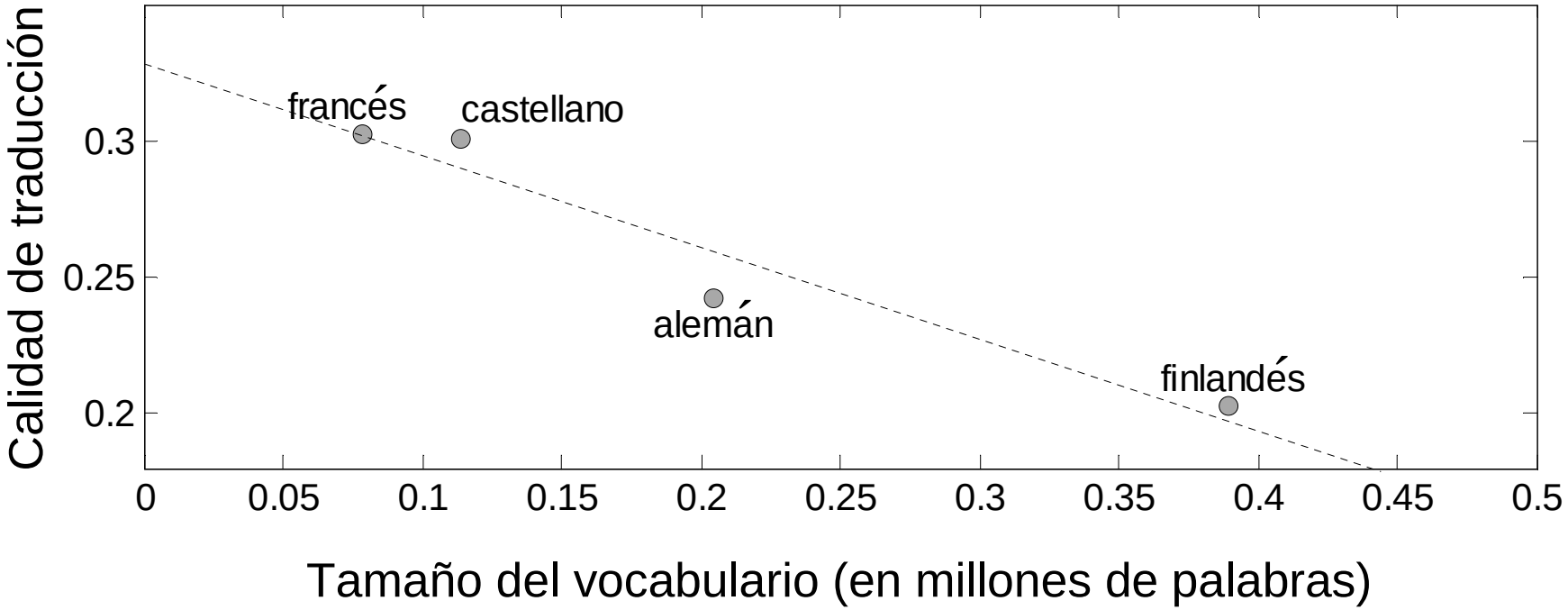
Retos futuros

Problemas específicos de la traducción automática

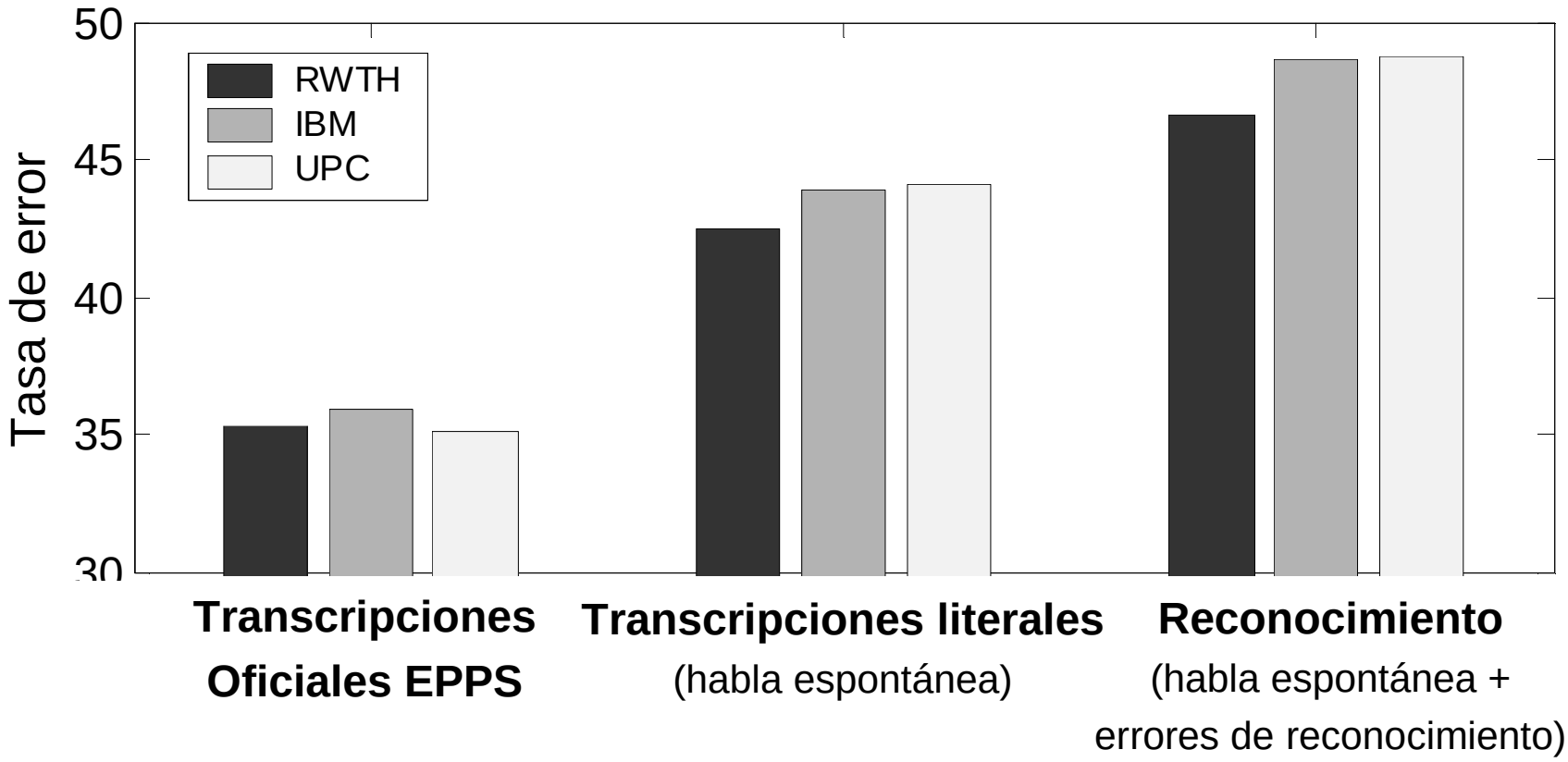
1.- Morfología: incide directamente sobre el tamaño del vocabulario, lo cual genera problemas de dispersión de los datos.

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

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



Efectos del lenguaje oral en traducción



El problema de las medidas de evaluación

1.- La evaluación humana es lenta y costosa.

2.- Las medidas de evaluación automáticas existentes:

- evalúan globalmente un conjunto de traducciones,
- no son confiables al evaluar traducciones individuales,
- dependen del conjunto de referencias disponible.

En los próximos años la investigación se centrará en:

1.- La incorporación de conocimiento lingüístico en el proceso de traducción automática estadística.

2.- El desarrollo de estrategias eficientes para abordar el problema de ordenamiento de palabras.

3.- El desarrollo de nuevos modelos para tomar en cuenta los efectos del habla espontánea.

4.- El desarrollo de medidas automáticas de evaluación más eficientes y confiables.

UNIVERSIDAD CATOLICA ANDRÉS BELLO
Facultad de Ingeniería
Escuela de Telecomunicaciones

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Traducción Automática Estadística

Rafael E. Banchs
Unversitat Politècnica de Catalunya

V Jornadas de Ingeniería de Telecomunicaciones – Universidad Católica Andrés Bello – Caracas, 10 de Mayo de 2007