C H I S P A – M T

Un Consorcio para el Desarrollo de Sistemas de Traducción Automática entre las Lenguas China y Castellana

Rafael E. Banchs
Josep M. Crego, Patrik Lambert, José B. Mariño
Univesitat Politècnica de Catalunya

1.- Creciente interés en tecnologías de traducción automática

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 ” + conference	*123.000*
“ machine translation ” + journal	*98.100*

** Julio 2005*

2.- Más proyectos y campañas internacionales de evaluación

- **GALE: Global Autonomous Language Environments**
  http://ciir.cs.umass.edu/research/nightingale.html
- **TC-STAR: Technology and Corpora for Speech to Speech Translation**
  http://www.tc-star.org
- **MANOS: Multilingual Application Network for Olympic Services**
  http://nlpr-web.ia.ac.cn/english/cip/english/project.htm
- **NIST: National Institute of Standards and Technology**
  http://www.nist.gov/speech/tests/mt/
- **IWSLT: International Workshop on Spoken Language Translation**
  http://www.is.cs.cmu.edu/iwslt2005/index.html

1.- Entre las lenguas más habladas del mundo

Lengua	*Hablantes nativos*
Chino	874.000.000
Castellano	332.000.000
Inglés	322.000.000
Hindi	291.500.000
Árabe	193.000.000
Bengalí	189.000.000

Lengua	*Incluyendo 2da Lengua*
Chino	1.051.000.000
Hindi	594.000.000
Inglés	510.000.000
Castellano	420.000.000
Ruso	255.000.000
Árabe	230.000.000

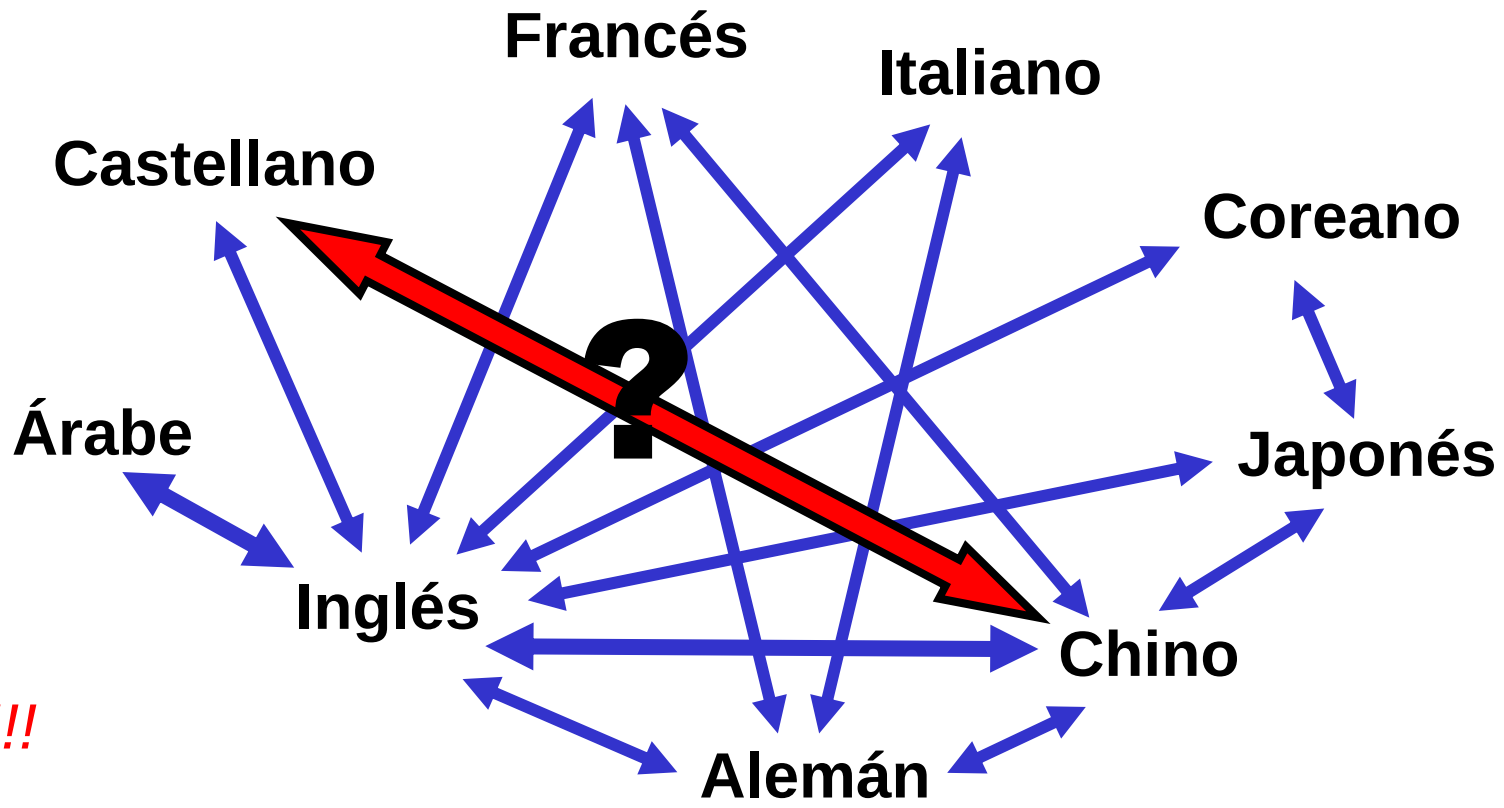
2.- Creciente interés en promover relaciones comerciales y tecnológicas entre países de habla Hispana y China:

- Programa bilateral Hispano-Chino de cooperación tecnológica.
- 1er foro empresarial de cooperación Hispano China de Málaga.
- Acuerdo de asociación estratégica entre España y China.
- Firma de contratos con empresas Chinas por parte de empresas españolas: Telefónica, Repsol, Indra, Gamesa...
- Adicionalmente: Chile, Argentina, México, Venezuela, etc...

3.- Poca (o ninguna) actividad en traducción automática en el par Chino-Castellano

- *Candide*
- *Verbmobil*
- *Nespole*
- *TC-star*
- *Gale*
- *Manos*
- *Otros...*

- *NINGUNO !!!*



3.- Poca (o ninguna) actividad en traducción automática en el par Chino-Castellano

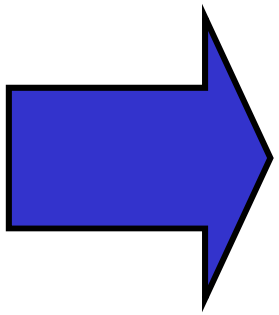
Búsqueda en *www.google.es**	Resultados
“ machine translation ” + Chinese + Spanish	*9.100*
“ machine translation ” + Chinese + English	*93.900*
“ machine translation ” + Spanish + English	*269.000*

** Marzo 2006*

Objetivos del Proyecto

1.- Investigación básica:

- Estudio de las características propias de la tarea.
- Adecuación de sistemas existentes a la tarea.
- Mejora del estado actual del arte en traducción automática. (aproximación estadística)



Sistema de traducción
Recursos bilingües (corpus, diccionarios...)
Recursos monolingües (corpus, modelos...)

2.- Investigación aplicada:

- Desarrollo de aplicaciones para la comunicación bilingüe.

Traducción asistida

Traducción automática

Mensajería SMS

Subtitulado

Navegación y búsqueda

Videoconferencia bilingüe

Traductor de e-mails

Asistente personal

Principales Retos Tecnológicos

- **Altas tasas de error y restricciones de dominio.**
- **Naturalezas del lenguaje oral y escrito.**
- **Morfología *(CASTELLANO).***
- **Reordenamiento de palabras y estructuras *(CHINO).***
- **Incorporación exitosa de conocimiento lingüístico.**

Cronograma de Actividades

1er año de actividades

- Desarrollo de un corpus bilingüe chino-castellano (1 Millón de oraciones)
- Desarrollo de un sistema de traducción “baseline”
- Primer prototipo de aplicación para la traducción asistida
- Primer prototipo de sistema de navegación bilingüe

2do año de actividades

- Depuración del corpus bilingüe chino-castellano
- Desarrollo de recursos lingüísticos generales para chino y castellano
- Versiones actualizadas de los sistemas de traducción y navegación bilingüe
- Primer prototipo de plataforma de traducción para mensajería electrónica

3er año de actividades

- Desarrollo de recursos de traducción para dominios específicos
- Versión comercial de los sistemas de traducción y navegación bilingüe

Cronograma de Actividades

3er año de actividades (continuación)

- Versión actualizada de plataforma de traducción para mensajería electrónica
- Primeros servicios relacionados a la plataforma de traducción

4to año de actividades

- Desarrollo de recursos para el reconocimiento y síntesis del habla
- Versión actualizada de plataforma de traducción para mensajería electrónica
- Desarrollo completo de servicios relacionados a la plataforma de traducción
- Versión actualizada de los sistemas de traducción y navegación bilingüe

5to año de actividades

- Desarrollo de sistemas de reconocimiento y síntesis del habla
- Evaluación y prueba de las aplicaciones desarrolladas
- Versión comercial de sistemas de traducción para mensajería electrónica
- Primer prototipo de herramientas de apoyo para video-conferencia bilingüe

Estimación de Recursos Requeridos

			YEAR 1		YEAR 2		YEAR 3		YEAR 4		YEAR 5	
	Description	**unit cos**	**amoun**	**total**	**amoun**	**total**	**amoun**	**total**	**amoun**	**total**	**amoun**	**total**
Resources	servers (research)	5000	3	15000	1	5000	0	0	0	0	0	0
	servers (development)	5000	2	10000	1	5000	1	5000	1	5000	0	0
	servers (web)	5000	1	5000	1	5000	0	0	1	5000	2	10000
	personal computers	1500	8	12000	2	3000	2	3000	0	0	0	0
	hardware update	2000	0	0	1	2000	2	4000	2	4000	2	4000
	software / licenses	20000	1	20000	1	20000	1	20000	1	20000	1	20000
	linguistic resources	10000	2	20000	2	20000	1	10000	1	10000	1	10000
Personnel	research / post-doc	25000	2	50000	3	75000	3	75000	3	75000	3	75000
	doctoral scholarships	15000	4	60000	6	90000	8	120000	8	120000	6	90000
	support staff	15000	3	45000	3	45000	4	60000	4	60000	4	60000
	software developers	20000	2	40000	3	60000	4	80000	5	100000	5	100000
	specific task personnel	15000	8	120000	4	60000	2	30000	2	30000	2	30000
Other expenses	conference attendance	1000	8	8000	10	10000	10	10000	10	10000	10	10000
	meetings and other visits	2000	3	6000	2	4000	2	4000	2	4000	3	6000
	other supplies & services	5000	2	10000	2	10000	3	15000	3	15000	3	15000
	Total for resources			82000		60000		42000		44000		44000
	Total for personnel			315000		330000		365000		385000		355000
	Total for other expenses			24000		24000		29000		29000		31000
	Total expenses			421000		414000		436000		458000		430000
	UPC overhead	30.00%		126300		124200		130800		137400		129000
	Total project cost			**547300**		**538200**		**566800**		**595400**		**559000**

Opciones de Financiamiento

1.- Financiación pública:

- Convocatorias de ayuda a la investigación
- Programas Universidad/Empresa: PROFIT, CENIT
- Programa bilateral de cooperación tecnológica CHINEKA

2.- Financiación privada:

- Convenio Empresa/Universidad (patrocinante único)
- Consorcio académico (múltiples patrocinantes)

Opciones de Financiamiento

1.- Financiación pública:

- Convocatorias de ayuda a la investigación
- Programas Universidad/Empresa: PROFIT, CENIT
- Programa bilateral de cooperación tecnológica CHINEKA

2.- Financiación privada:

- Convenio Empresa/Universidad (patrocinante único)
- Consorcio académico (múltiples patrocinantes)

Consorcio Académico

Principal Socio Investigador

Centro TALP (http://www.talp.upc.edu/talp/)

Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla

- Grupo de procesamiento del lenguaje natural (GPLN)
- Grupo de procesado de la voz (VEU)

***** Dedicado a proyectos de investigación básica y aplicada

***** Participación en proyectos españoles y europeos

***** Cooperación con instituciones públicas y empresas

***** Más de 15 años de experiencia en tecnologías del lenguaje

Experiencia en Traducción Automática

- **Participación en proyectos españoles y europeos**
  (ALIADO, FAME, LC-STAR, TC-STAR-P, TC-STAR)
- **Participación en campañas internacionales de evaluación**
  (IWSLT04, ACL05, TC-STAR05, IWSLT05, TC-STAR06)
- **Sistema propio con prestaciones en el estado del arte**
  (D5: SLT Progress Report, www.tc-star.org/)
- **Experiencia en varios pares de lenguas**
  (CA<->ES, ES<->EN, FR<->EN, ZH->EN, AR->EN, GE->EN, FI->EN)
- **Relación con otros centros de investigación:** CMU, UKA, RWTH
- **Demostrador en línea castellano-catalán:** ***http://www.n-ii.org***

C H I S P A – M T

Un Consorcio para el Desarrollo de Sistemas de Traducción Automática entre las Lenguas China y Castellana

Rafael E. Banchs
Josep M. Crego, Patrik Lambert, José B. Mariño
Univesitat Politècnica de Catalunya