TC-STAR
A Speech to to Speech Translation project

Gianni Lazzari

IWSLT 2006

Kyoto November 27

Summary

- Why Speech to Speech Translation (SST) ?
- TC-STAR project
- Second Evaluation in TC-STAR
  - tasks and conditions, data, participants, results
  - technologies evaluated : ASR, SLT, TTS
  - automatic and human evaluation
- Conclusions

Why SST ?

- **To let people communicate**
  - Telephone conversation
  - Face to face

- **To let people understand news and content produced in foreign languages:**
  - Internet, Conferences, Multimedia Documents, Broadcast, Lectures..
    - **Off-line**
    - **Simultaneously.**

SST projects in the last 20 years

- **Pioneers**
  - C-STAR
  - IBM (statistical machine translation)
- **Demonstration oriented and limited domain**
  - C-STAR II – VERBMOBIL - NESPOLE! - BABYLON – DIGITAL OLIMPICS
- **Technology oriented and limited domain**
  - **C-STAR III (IWSLT)**
- **Technology oriented and unlimited domain**
  - TC-STAR
  - STR-DUST
  - GALE

TC-STAR

Transcription and Translation of broadcast news, speeches and interviews

Vocal access

Simultaneous Translation

Web access

TC-STAR

TC-STAR Project focuses on advanced research in key technologies for speech to speech translation (SST):

- ***speech recognition (ASR);***
- ***spoken language translation (SLT);***
- ***speech synthesis (TTS).***

- Start: April 2004
- End: March 2007
- Grant: 11 M. Euro

Objectives

The objective of the project is to reach a breakthrough in SST research in order to minimize the gap between human and machine performance. This objective will be pursued through:

- ***the development of new algorithms and methods;***

- ***the realization of a SST technology evaluation infrastructure to measure progress via competitive evaluation;***

- ***the integration of the SST technology components helps establishing de-facto standards for SST systems*.**

PARTNERS


RWTH AACHEN
RHEINISCH-WESTFÄLISCHE TECHNISCHE HOCHSCHULE AACHEN
SPEX
UNIVERSITAT POLITÈCNICA DE CATALUNYA
UPC
IBM
ELDA
Universität Karlsruhe
SONY
NOKIA
CONNECTING PEOPLE
ITC irst
SIEMENS

Application Scenario

– A selection of unconstrained conversational speech domains:

- ***Broadcast news***
- ***European Parliament Speeches***

– A few languages important for Europe society and economy:

- ***European Accented English***
- ***European Spanish***
- ***Mandarin***

European Parliament Scenario

- Highly scalable scenario overall Europe
  - 380 language pairs with 20 official languages
- Highly motivated by accessibility and inclusion. A huge amount of info is not accessible!
- Recordings from *Europe by Satellite (EbS)*
  - Source language (speakers)
  - Target languages (interpreters)
- Texts from EU translation service

European Parliament audio data training October 2006 status

detail			acoustic amount [h] English EPPS	Spanish EPPS	Spanish PARL
total amount of recordings			289.3	286.1	44.2
transcribed recordings			102.1	98.9	44.2
untranscribed 2005/02-2005/05			75.2	75.2	–
untranscribed 2005/12-2006/05			112.0	112.0	–
transcribed speeches			91.6	61.9	38.4
male	interpreter	native	40.4	22.1	–
		non-native	0.9	1.8	–
	politician	native	11.0	8.8	27.3
		non-native	6.3	0.3	1.4
female	interpreter	native	26.0	24.4	–
		non-native	3.4	3.0	–
	politician	native	2.9	1.3	9.7
		non-native	0.6	0.2	–

DATA BROADCASTED BY EBS

Workplan

- First Evaluation Campaign (*internal*) & workshop: ***Trento April 2005***
- Second Evaluation Campaign (*open*) & workshop: ***Barcelona 2006***
- Third Evaluation Campaign (*open with Infrastructure*) & workshop: ***Aachen 2007***

- Showcase of SST results

Second evaluation campaign
February 1 - March 15 2006

.....

to measure progress in the second year of the project in the *three technologies* and in *the integration of the components*.

Workshop
Barcelona June 2006

Challenges for second evaluation

- Fully automatic evaluation
  - *without manual segmentation*
- Parliament data: politicians only
- Additional task for portability:
  - *Cortes data for Spanish to English*
- Open evaluation & Comparison with Systran
- Evaluation procedure:
  - *evaluation measures with missing segmentation*
  - *human evaluation and end-to-end evaluation*
- System combination
- General improvements in technology

Participants

	Automatic Speech Recogntition			Spoken Language Translation			Text To Speech		
	EN	ES	ZH	EN→ES	ES→EN	ZH→EN	EN	ES	ZH
IBM	X	X		X	X		X	X	X
ITC-irst	X	X		X	X	X			
LIMSI	X	X	X		X				
NOKIA	X						X		X
RWTH	X	X		X	X	X			
SIEMENS							X	X	
SONY									
UKA	X		X	X	X	X			
UPC				X	X		X	X	
ATT							X	X	
CAS									X
DFKI				X	X				
ICT						X			
NLPR						X			
NRC						X			
U. Endinburgh				X	X				
Univ. Dresden							X		
Univ. Muncih							X		
U. Vigo		X							
U. Washington				X	X				

Table 1 Participants in the Second TC-STAR Evaluation Campaign

Overview of the Campaign

- Evaluated Technologies: 3 out of 3
  - ASR - SLT -TTS
- Schedule: from February 1 2006 to March 15 2006
- Participants
  - 8 for ASR: 7 En, 6 Es, 1 Zh; 1 external
    33 submissions (22 En, 10 Es, 1 Zh)
  - 13 for SLT: 8 EnEs, 9 EsEn, 6 ZhEn; 6_external
    116 submissions (38 EnEs, 45 EsEn, 33 ZhEn)
  - 10 for TTS: 4 external
    61 submissions (26 En, 26 Es, 9 Zh)

Evaluation Tasks

2 tasks

- **PARLIAMENTARY SPEECHES**
  - English (En) and Spanish (Es) from the European Parliament Plenary Sessions
  - Spanish from the Cortes
- **BROADCAST NEWS** Mandarin Chinese (Zh), Broadcast News from Voice of America (partly supplied by LDC)

Evaluation data

- In order to chain ASR SLT and TTS components, evaluation tasks have been designed to use *common data sets* of raw data and conditions

ASR Tasks

- 2 Tasks
  - **PARLIAMENT**:
    - EPPS English 3 hours
      ~34 K words
    - EPPS Spanish 3 hours
      ~32 K words
    - CORTES Spanish 3hours
      ~32 K words
  - **BN**
    - Zh : 3 hours of VoA recorded in Dec 1998
      ~42 K characters

- 3 Conditions
  - **Restricted** training condition (ie TC-Star data)
  - **Public data condition (**ie data available through ELDA and LDC**)**
  - **Open condition** (any data before May 31 2005)

Language Resources for ASR

Training	
EPPS	101 h manual transcripts + 75h non transcribed from sessions recorded May 04 – May 05 → produced by RWTH (En) and UPC (Es)
CORTES	**40 h manual transcripts -> UPC**
VOA	publicly available sources + VOA 1998 available at LDC without Dec. 1998 (audio + LDC transcripts)
Development	
EPPS	6h audio + manual transcripts (sessions June, July 2005) → produced by ELDA (En+Es)
CORTES	3h audio + manual transcripts (sessions Dec 2004) → produced by UPC
VOA	3h audio + manual transcripts (1-11 Dec 1998) → transcripts by ELDA
Test	
EPPS	6h audio + manual transcripts (sessions 15-18 Nov 2004) → produced by ELDA
CORTES	3h audio + manual transcripts (sessions Nov 2005) → produced by ELDA
VOA	3h audio + resegmented manual transcripts (22-25 Dec 1998) → transcripts by ELDA

Language Reference for public condition

Language	Reference	Amount
Chinese	Mandarin 1997 BN (Hub4-NE) LDC98S73 (audio) & LDC98T24 (transcr)	~30h
	Mandarin 2001 Call (Hub5) LDC98S69, LDC98T26 (transcr)	~40h
	Mandarin TDT2 LDC2001S93 & LDC2001T57 (transcr)	
	Mandarin TDT3 LDC2001S95 & LDC2001T58	
	Mandarin Chinese News Text LDC95T13	250M words
	Mandarin CALLHOME LDC96S34, LDC96T16 (transcr)	
	Chinese Gigaword LDC2003T09	1.1G words
	Hong Kong News Parallel Text LDC2000T46 (Zh/En)	18147 articles
Spanish	EPPS_SP (text): Apr 1996 - May 2005	>36M words
	TC-STAR_P Spanish BN	10h transcribed
	Spanish LDC 1997, BN speech (Hub4-NE), LDC98S74	
	Spanish LDC CallHome, LDC96S35	
English	EPPS_EN (text): Apr 1996 - May 2005	>36M words
	TC-STAR_P English BN	10h transcribed
	English LDC 1995 (CSR-IV Hub 4 Marketplace LDC96S31), 1996, 1997, official NIST Hub4 training sets, LDC97S44 and LDC98S71, USC Marketplace Broadcast News Speech (LDC99S82)	
	English LDC TDT2 and TDT3 data with closed-captions, about 2000h, LDC99S84 and LDC2001S94	
	English LDC Switchboard 1, 2-I, 2-II, 2-III, LDC97S62, LDC98S75, LDC99S79	
	English LDC Callhome, LDC97S42, LDC2004S05, LDC2004S09	
	English LDC Meeting corpora, ICSI LDC2004S02, ISL LDC2004S05, NIST LDC2004S09	

Table 3 Public condition training resources

English Results (22 submissions)

Case insensitive scoring

System	2005 Systems Eval05 data Open/Public	2005 Systems Eval05 data Restricted	2006 Systems Eval06 data Open/Public	2006 Systems Eval06 data Restricted	Gain
TCStar	9.5	-	6.9	-	-27%
IBM	11.6	12.3	8.8		-25%
IRST	-	13.4	11.0	-	-18%
LIMSI	10.6	11.2	8.2	-	-23%
NOKIA	24.6	-	18.3	-	-26%
RWTH	-	14.1	-	10.2	-28%
SONY	50.0	-	-	35.9	-28%
UKA	14.0	-	10.0		-29%

Spanish Results

Case insensitive scoring, restricted training

Systems	*2005 Systems Eval05 data EPPS*	*2006 Systems Eval 06 data EPPS*	*Cortes*	EPPS+Cortes	*Gain*
TCStar	10.1	6.2	9.8	8.1	-39%
IBM	12.2	8.3	12.5	10.6	-30%
IRST	13.7	9.7	11.0	13.5	-29%
LIMSI	11.5	7.8	13.3	10.7	-32%
RWTH	12.7	8.0	12.1	10.2	-37%
UVIGO	-	20.1	35.7	28.4	-

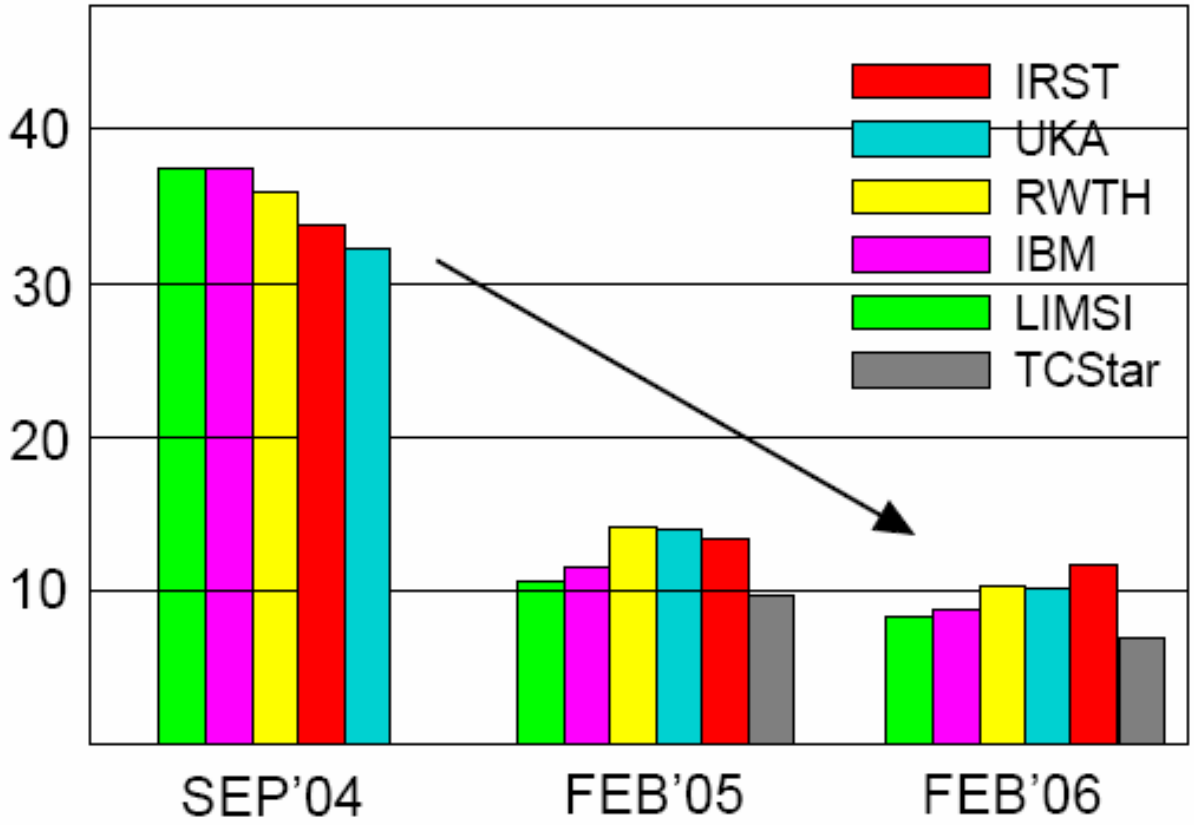
ASR Chinese results

Cross adaptation

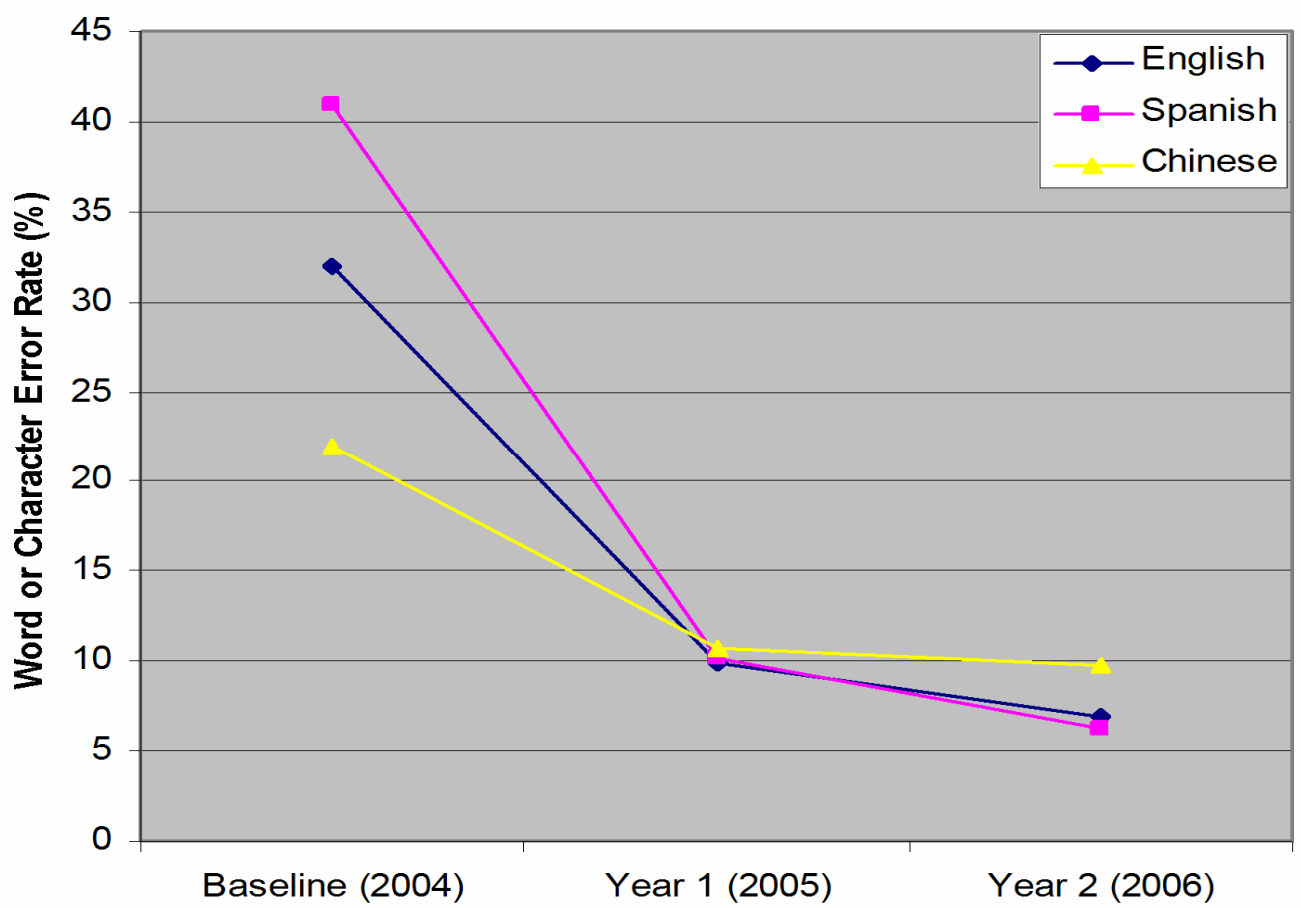
1. Run UKA system → HYP1
2. Adapt LIMSI acoustic models using HYP1
3. Run LIMSI system → HYP2

System	*2005 System Eval05 data Dev data*	*2006 System Eval 06 data Eval data*
LIMSI/UKA	10.7	9.8

Progress Summary on English EPPS



EPPS English: 32.0 → 10.6 (-67%), Rover: 9.5 → 6.9% (-27%)


45
40
35
30
25
20
15
10
5
0
Word or Character Error Rate (%)
English
Spanish
Chinese
Baseline (2004)
Year 1 (2005)
Year 2 (2006)

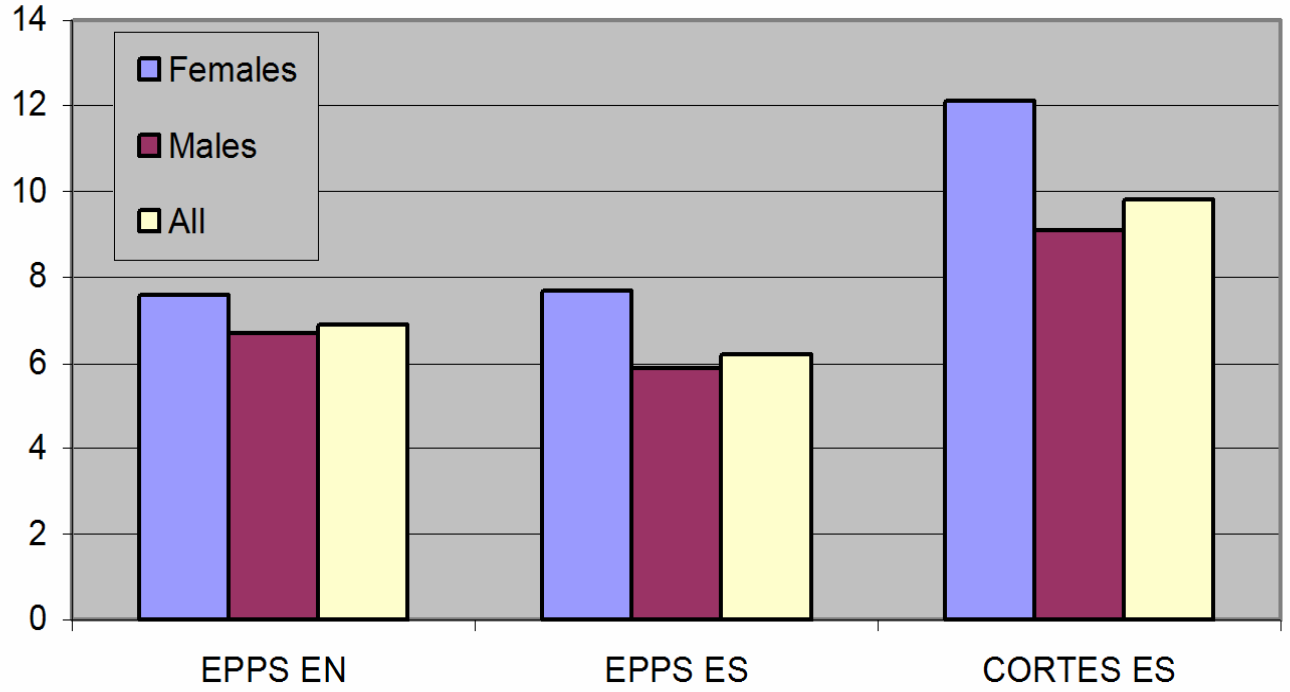


Figure 1 TC-STAR system performance for male and female speakers

Most common substitution

Confusion pairs for English	Confusion pairs for Spanish	Confusion pairs for Mandarin
a / the	las / la	她 / 他
and / in	Del / el	了 / 的
the / a	el / del	它 / 他
(%hpwesitation) / and	(%hesitation) / de	的 / 地
that / the	(%hesitation) / que	利 / 力
the / that	del / de	是 / 时
or / all	el / al	作 / 做
too / to	(%hesitation) / en	地 / 的
been / being	al / el	呢 / 的
had / have	de / del	是 / 使

Table 8: Top ten substitution errors for the English and Spanish EPPS task and the Mandarin VOA task

Main Achievements

- Best word error rate on English and Spanish EPPS are 8.2% and 7.8%
  - *most errors are substitutions*
  - *better system performance for male (only 25% of female data)*
  - *worse performance by non native speakers*
- System combination: 6.9% for English and 8.1% for Spanish(EPPS+CORTES)
- Almost 30% compared to be best systems in the TC-STAR Mar'05 evaluation
- Automation of the segmentation step needed for SLT-MT
- Production of transcriptions, enriched segmentation, casing, punctuation.

SLT Tasks

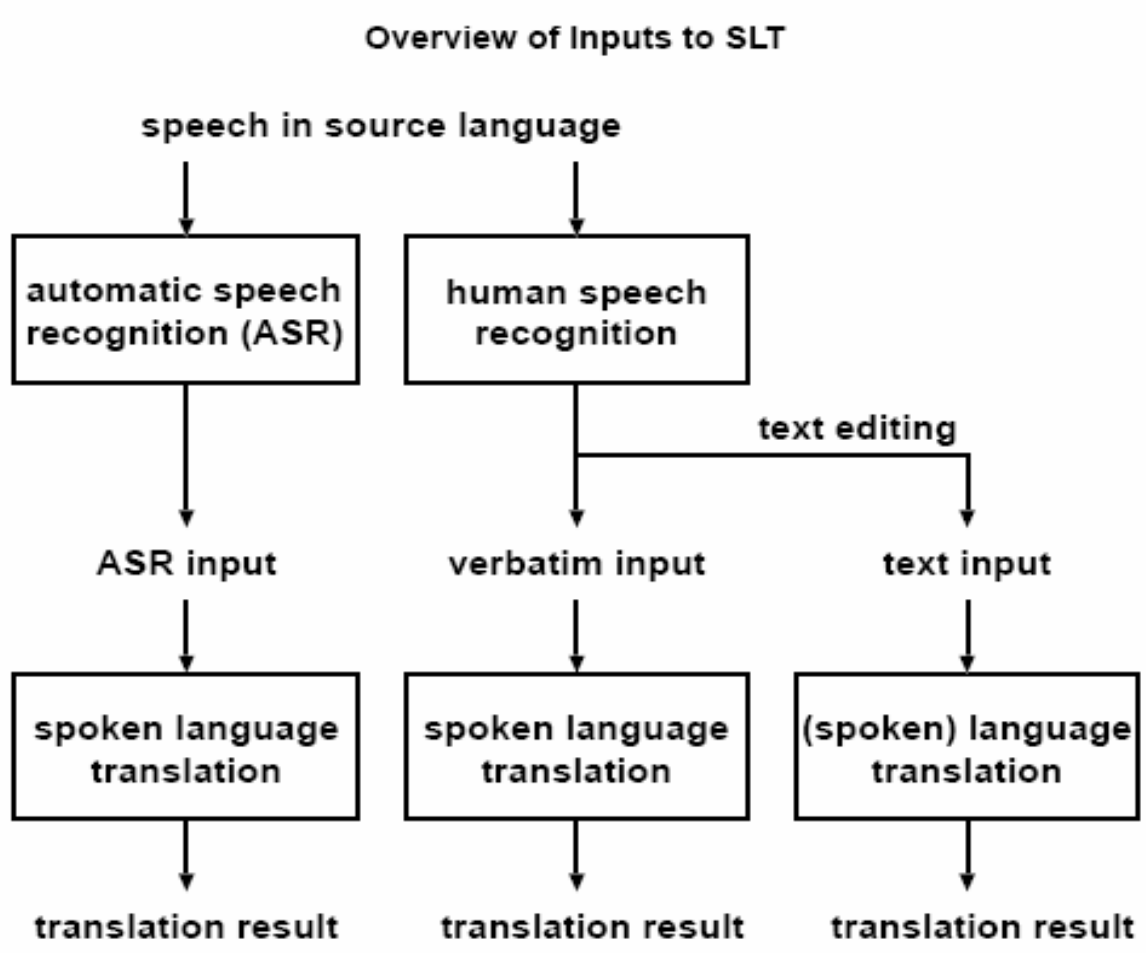
Four evaluation tasks:

- English-Spanish: EPPS (European Parliament Plenary Sessions)
- Spanish-English: EPPS (European Parliament Plenary Sessions)
- Spanish-English: CORTES (Spanish Parliament) *to study portability*
- Chinese-English: BC News *to study language pairs with different structure and comparison with US projects*

Three input conditions: *to study the effect of ASR errors and spontaneous speech*

- ASR input:
  - identical input to ALL systems!
  - automatic sentence segmentation
- verbatim transcriptions
- text

Inputs to SLT



SLT training conditions

Primary:

- English→Spanish and Spanish→ English: EPPS data produced in TC-STAR
- Chinese→English: LDC data listed in the training data table

Aim: *strict comparison of the systems*

Secondary:

- any pubblicly available data before the cut-off date may 31 2005

Aim: *comparison of the systems without data constraints*

SLT Training Data Set

Direction	Data
Zh->En	FBIS Multilanguage Texts
	UN Chinese English Parallel Text Version 2
	Hong Kong Parallel Text
	English Translation of Chinese Treebank
	Xinhua Chinese-English Parallel News Text Version 1.0 beta 2
	Chinese English Translation Lexicon version 3.0
	Chinese-English Name Entity Lists version 1.0 beta
	Chinese English News Magazine Parallel Text
	Multiple-Translation Chinese (MTC) Corpus
	Multiple Translation Chinese (MTC) Part 2
	Multiple Translation Chinese (MTC) Part 3
	Chinese News Translation Text Part 1
	Chinese Treebank 5.0
	Chinese Treebank English Parallel Corpus
Es->En	EPPS Spanish verbatim transcriptions May 2004 - Jan 2005
	EPPS Spanish and English Final Text Edition May 2004- Jan 2005
	EPPS Spanish Final Text Edition April 1996 to Jan 2005
En->Es	EPPS English verbatim transcriptions May 2004- Jan 2005
	EPPS English to Spanish Final Text Edition May 2004- Jan 2005
	EPPS English Final Text Edition April 1996 to Jan 2005

SLT Development Data Set

Direction	Data	Epoch
Zh->En	VOA Verbatim transcriptions with 2 references translations	From December 14, 1998 to December 16, 1998
	VOA ASR transcriptions	
Es->En	EPPS verbatim transcriptions with 2 reference translations	From June 6, 2005 to July 7, 2005
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	
	EPPS verbatim transcriptions with 2 reference translations	December 1 & 2, 2004
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	
En->Es	EPPS verbatim transcriptions with 2 reference translations	From June 6, 2005 to June 9, 2005
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	

SLT Test Data Set

Direction	Data	Epoch
Zh->En	VOA Verbatim transcriptions with 2 references translations	From December 23, 1998 to December 25, 1998
	VOA ASR transcriptions	
Es->En	EPPS verbatim transcriptions with 2 reference translations	From September 5, 2005 to November 17, 2005
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	
	EPPS verbatim transcriptions with 2 reference translations	November 24, 2005
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	
En->Es	EPPS verbatim transcriptions with 2 reference translations	From September 7, 2005 to September 26, 2005
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	

SLT Data Set format

- **22 data sets.**
- For each set there are:
  - *The data to be translated in the source language*, organized in documents and segments, except the ASR input which is in CTM format
  - *Two reference translations of the source data*, issued by professional translators, also organized in documents and segments.
  - *Several candidate translations* produced by the participants in the evaluation, following the same format of the source and reference sets.

Validation of Language Resources

- Reference translations of dev and test sets for all the three translation directions were validated on a statistical based with the following penalty scheme:

Error	Penalty points
Syntactical	4 points
Deviation from guidelines	3 points
Lexical	2 points
Poor usage	1 points
Punctuation or spelling errors	0.5 (with a maximum of 10)

Table 12 LRs translation errors

SLT Participants

TC-Star participants:

- **IBM: IBM Research Yorktown Heights, USA**
- **ITC-irst: ITC-irst Trento, Italy**
- **LIMSI: LIMSI-CNRS Paris, France**
- **RWTH: RWTH Aachen University, Germany**
- **UKA: University of Karlsruhe (jointly with CMU), Germany**
- **UPC: Universidad Politecnica de Catalunya, Spain**

SLT External Participants

Spanish→English:

- DFKI: German Center for Artificial Intelligence, Saarbrücken, Germany
- UED: University of Edinburgh, Scotland, UK
- UWA: University of Washington, Seattle, USA

Chinese→English:

- ICT: Institute of Computing Technology, Beijing, China
- NLPR: National Laboratory for Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
- NRC: National Research Council, Ottawa, Canada

Moreover a off the shelf Systran Product has been evaluated by ELDA

Submissions

- Total number 116
- 38 for En→SP
- 45 for SP→EN
- 33 for Zh→EN

Submissions

Site	En→Es			Es->En			Zh->En	
	ASR	FTE	Verbatim	ASR	FTE	Verbatim	ASR	Verbatim
IBM	4P	3P	3P	4P	4P	4P		
ITC-irst	1P + 1S	1P + 1S	1P	1P + 2S	1P + 2S	1P + 2S	1P + 3S	1P + 3S
LIMSI				1P		1P		
RWTH	2P	2P	3P	2P	2P	3P	3P	3P
UKA	1P	1P	1P	1P	1P	1P	2P	4P
UPC	1P	2P + 1S	1P	1P	1P + 1S	1P		
DFKI		1P			1P	1P		
ICT							3P	6P
NLPR							1S	1S
NRC								1P + 1S
UED		1P			1P			
UW		2P	2P		2P	2P		

Table 14 List of submissions in the Second TC-STAR Evaluation Campaign .Condition types: P: Primary; S: Secondary

Evaluation Results

- The same ASR input used for all the systems:
  - TC-STAR ROVER English and Spanish
  - LIMSI/UKA for Mandarin
- Case information was used by evaluation metrics
- Punctuation marks presents in all the inputs, but Mandarin.

Human Evaluation

- English to Spanish only

- Segment produced by ASR, Verbatim, FTE and all their reference translations evaluated in relation to *adequacy and fluency*
- *Adequacy:* target segments compared to reference segments
- *Fluency:* quality of grammar evaluated

Human Evaluation

- Evaluators assess all the segments first accordingly to Fluency and then to Adequacy, so that:
  - Both types of measures are done independently
  - Each evaluator assesses both for a certain number of segments

Human Evaluation

- Evaluation of *Fluency*:
  - Answer to the question “Is the text written in good Spanish?”
  - 5 points scale, where only extreme marks defined:

    1= Not Understandable 5= Perfect
- Evaluation of *Adequacy:*
  - Answer to the question:” How much of the meaning expressed in the reference translation is also expressed in the target translation?”
  - 5 points scale, where only extreme marks defined:

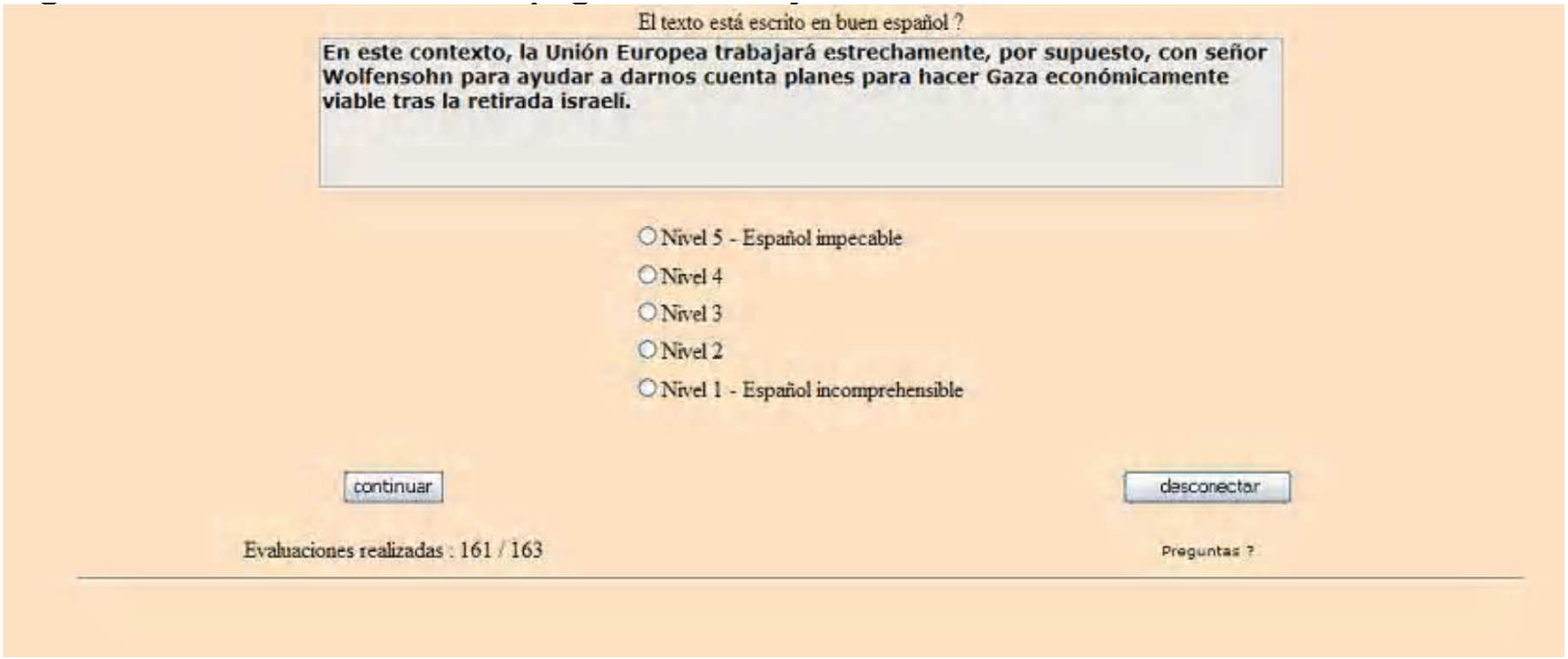
    1= Nothing in common 5= All the meanings

Human Evaluation

- 2 evaluations per segment by different evaluators
- Evaluators are native speakers of the target language up to University level
- No knowledge of the source language required
- Segments presented randomly

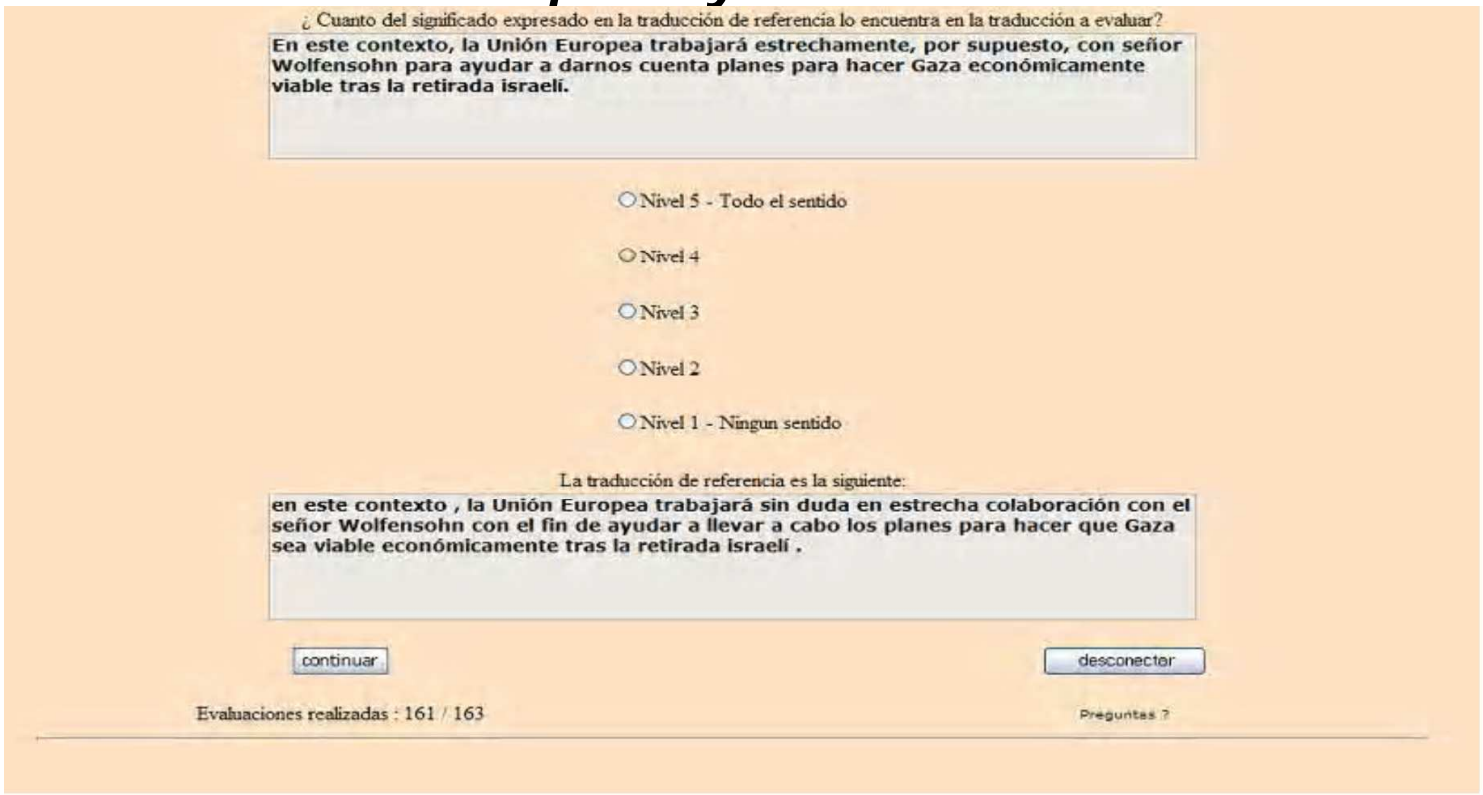
Human Evaluation

- On line evaluation based on a Web interface : *Fluency*



Human Evaluation

- On line evaluation based on a Web interface : *Adequacy*



Human Evaluation

- Figures about the human evaluation

Number of evaluators	number of evaluation / segment	Task	Number of segments	Number of translation / segment	Total number of evaluations	#Evaluation segments / Evaluator
125	2	FTE	392	11	8,624	162.88
		Verbatim	388	9	6,984	
		ASR	396	6	4,752	

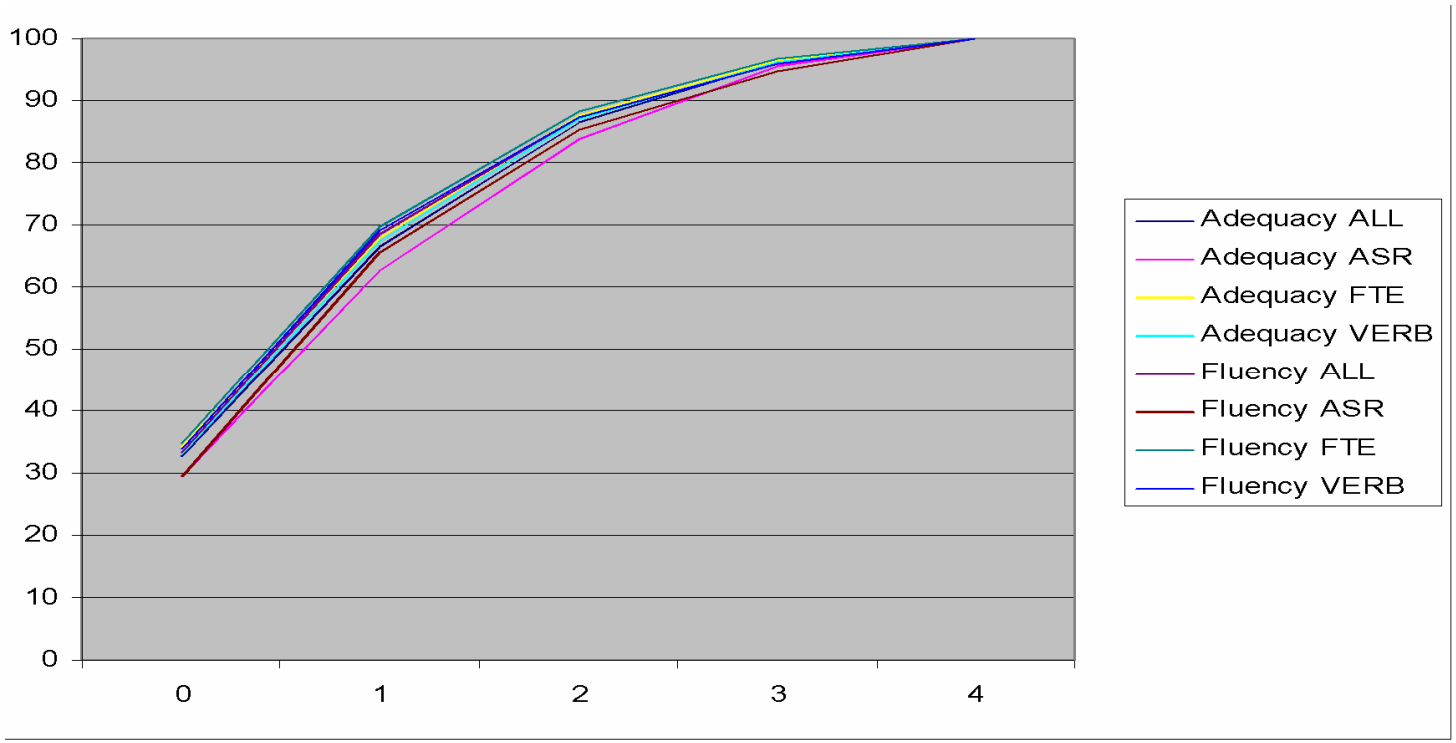
Human Evaluation

- Evaluator agreement:
  - Total agreement between evaluators rather good: about un third of segment obtained identical evaluations within the two evaluators

	FTE + Verb. + ASR	FTE	Verbatim	ASR
Fluency	33.16	34.74	33.85	29.29
Adequacy	32.64	34.23	32.82	29.50

Consistency of the score

Agreement between the first and the second scores for all the segments computed as a function of the difference between the first and the second scores: >30% same score ; 65% diff =1



Evaluation Results FTE Task

SYSTEM	Fluency 5 : good 1 : bad	Adequacy 5 : good 1 : bad	Ranking Fluency	Ranking Adequacy
Human Reference	*4.56*	*4.44*	*1*	*1*
UED	3.63	3.79	2	2
RWTH	3.58	3.74	3	3
IBM	3.50	3.60	4	8
UPC	3.48	3.68	5	5
IRST	3.46	3.67	6	6
ROVER	3.46	3.72	6	4
UW	3.40	3.62	8	7
DFKI	3.31	3.53	9	9
UKA	3.17	3.49	10	10
SYSTRAN[4]	*2.46*	*2.93*	*11*	*11*

Table 19: Human scoring and ranking for the FTE task

Ranking by each evaluator

Fluency	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	Mean	Rank
Human Reference	*88*	*14*	*1*	*5*	*6*	*2*	*4*	*3*	*1*	*1*	*0*	*2.02*	*1*
UED	13	15	23	15	18	14	6	7	9	4	1	4.61	2
RWTH	4	14	12	14	18	16	12	16	8	5	6	5.68	3
IBM	2	13	15	20	12	12	11	13	12	9	6	5.84	4
IRST	6	15	14	9	12	11	17	20	5	6	10	5.87	5
ROVER	1	15	16	13	13	14	15	9	9	15	5	5.94	6
UW	4	10	13	12	12	12	9	16	17	15	5	6.34	7
UPC	1	9	11	13	14	14	14	15	20	14	0	6.37	8
DFKI	3	8	9	11	9	13	21	12	15	19	5	6.69	9
UKA	2	6	11	9	7	12	8	8	17	27	18	7.45	10
SYSTRAN	*1*	*6*	*0*	*4*	*4*	*5*	*8*	*6*	*12*	*10*	*69*	*9.20*	*11*

Table 20: Mean rank for FTE fluency

Adequacy	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	Mean	Rank
Human Reference	*68*	*15*	*5*	*6*	*5*	*9*	*6*	*4*	*3*	*1*	*3*	*2.88*	*1*
UED	13	16	12	18	17	21	7	8	4	6	3	4.85	2
RWTH	10	14	11	16	15	7	9	12	12	10	9	5.74	3
ROVER	5	9	19	16	14	9	16	9	13	10	5	5.79	4
IRST	6	12	14	17	8	17	9	13	14	13	2	5.82	5
UPC	5	11	13	11	16	5	20	19	12	5	8	6.06	6
UW	4	14	17	7	9	12	10	14	17	8	13	6.29	7
IBM	2	7	11	12	15	16	16	13	15	9	9	6.46	8
DFKI	8	6	9	10	10	13	13	11	21	20	4	6.62	9
UKA	3	16	7	7	9	11	11	17	7	29	8	6.79	10
SYSTRAN	*1*	*5*	*7*	*5*	*7*	*5*	*8*	*5*	*7*	*14*	*61*	*8.70*	*11*

Table 21: Mean rank for FTE adequacy

Evaluation Results
Verbatim Task

SYSTEM	Fluency 5 : good 1 : bad	Adequacy 5 : good 1 : bad	Ranking Fluency	Ranking Adequacy
Human Reference	*4.31*	*4.31*	*1*	*1*
UPC	3.39	3.54	2	4
RWTH	3.38	3.55	3	2
IBM	3.35	3.51	4	6
IRST	3.35	3.54	4	4
ROVER	3.32	3.55	6	2
UW	3.14	3.43	7	7
UKA	3.07	3.36	8	8
SYSTRAN	*2.34*	*2.77*	*9*	*9*

Table 22: Human scoring and ranking for the Verbatim task

Ranking by each evaluator
Verbatim Task

Fluency	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	Mean	Rank
Human Reference	*87*	*18*	*6*	*2*	*2*	*4*	*4*	*2*	*0*	*1.82*	*1*
UPC	12	16	16	30	12	8	15	11	5	4.46	2
IBM	2	16	17	29	19	14	12	9	7	4.79	3
ROVER	7	21	20	11	10	16	17	16	7	4.87	4
IRST	2	15	23	12	20	24	10	15	4	4.95	5
RWTH	2	20	11	14	29	18	18	6	7	4.97	6
UW	6	8	12	11	14	17	26	24	7	5.69	7
UKA	4	9	12	11	15	13	14	30	17	5.97	8
SYSTRAN	*3*	*2*	*8*	*5*	*4*	*11*	*9*	*12*	*71*	*7.48*	*9*

Table 23: Mean rank for Verbatim fluency

Adequacy	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	Mean	Rank
Human Reference	*71*	*20*	*9*	*12*	*3*	*5*	*0*	*5*	*0*	*2.168*	*1*
ROVER	9	23	18	12	19	11	12	15	6	4.608	2
UPC	7	17	17	21	15	15	15	12	6	4.768	3
UW	12	11	21	7	13	18	14	20	9	5.096	4
RWTH	4	11	16	23	18	14	20	10	9	5.128	5
IRST	6	16	11	12	18	24	17	12	9	5.192	6
IBM	8	13	9	18	19	18	15	15	10	5.208	7
UKA	6	10	18	17	9	16	16	17	16	5.448	8
SYSTRAN	*2*	*4*	*6*	*3*	*11*	*4*	*16*	*19*	*60*	*7.384*	*9*

Table 24: Mean rank for Verbatim adequacy

Evaluation Results
ASR Task

SYSTEM	Fluency 5 : good 1 : bad	Adequacy 5 : good 1 : bad	Ranking Fluency	Ranking Adequacy
RWTH	3.06	3.13	1	1
IBM	3.04	3.05	2	4
UPC	3.04	3.09	2	2
IRST	2.99	3.09	4	2
UKA	2.84	2.97	5	5
SYSTRAN	*2.09*	*2.33*	*6*	*6*

Table 25: Human scoring and ranking for the ASR task

Ranking by each evaluator
ASR Task

Fluency	1st	2nd	3rd	4th	5th	6th	Mean	Rank
RWTH	37	25	27	20	12	4	2.66	1
IBM	26	30	21	25	15	8	2.98	2
IRST	17	27	26	27	13	15	3.30	3
UPC	21	13	29	26	26	10	3.42	4
UKA	13	22	15	17	47	11	3.77	5
SYSTRAN	*11*	*8*	*7*	*10*	*12*	*77*	*4.88*	*6*

Table 26: Mean rank for ASR fluency

Adequacy	1st	2nd	3rd	4th	5th	6th	Mean	Rank
RWTH	31	28	25	18	17	6	2.84	1
UPC	21	28	24	14	27	11	3.25	2
IBM	25	21	20	27	19	13	3.26	3
IRST	23	22	19	29	23	9	3.27	4
UKA	17	20	24	23	28	13	3.51	5
SYSTRAN	*8*	*6*	*13*	*14*	*11*	*73*	*4.86*	*6*

Table 27: Mean rank for ASR adequacy

Summary

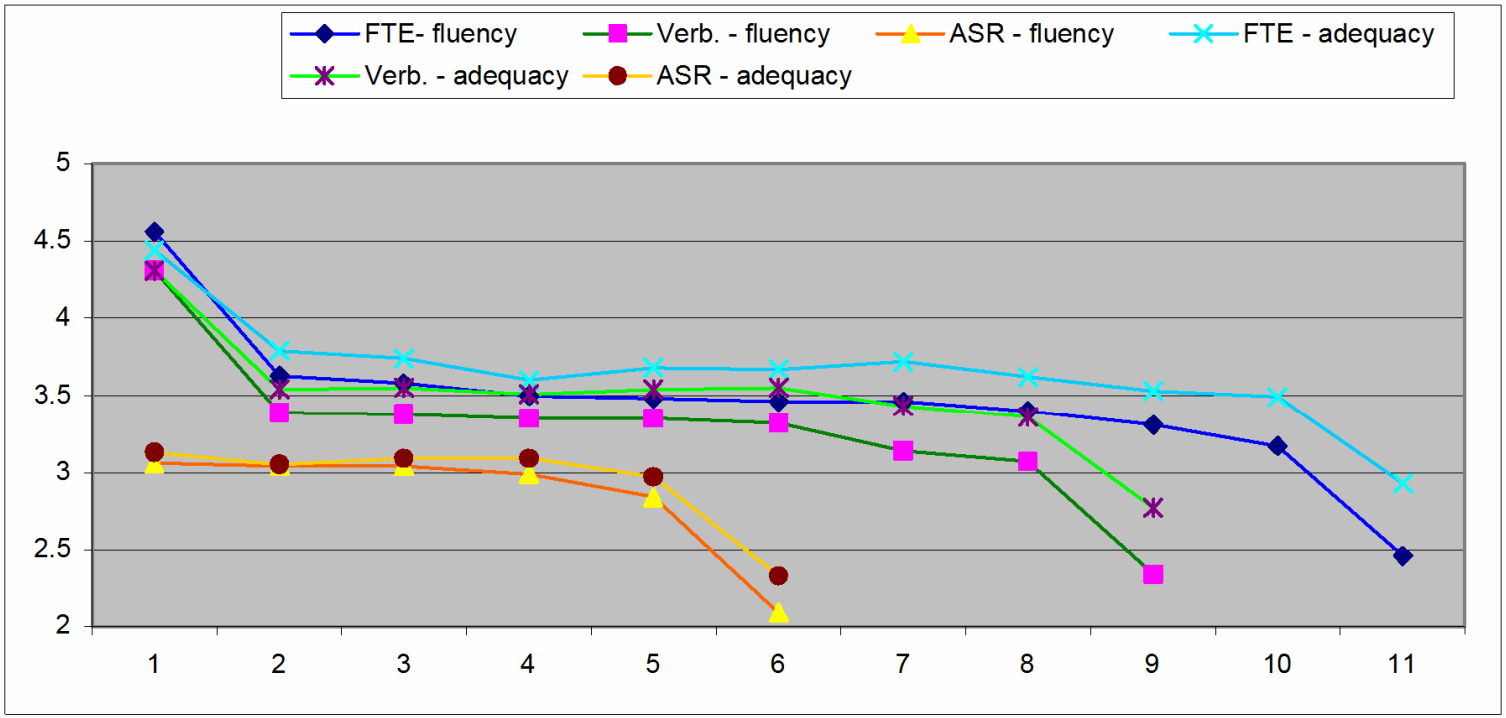


Figure 6: Differences between FTE, Verb. and ASR scores

Summary

Task	Site	Score fluency ranking	Mean rank fluency ranking	Score adequacy ranking	Mean rank adequacy ranking
FTE	*Human Reference*	*1*	*1*	*1*	*1*
	UED	2	2	2	2
	RWTH	3	3	3	3
	IBM	4	4	8	8
	UPC	5	8	5	6
	ROVER	6	6	4	4
	IRST	6	5	6	5
	UW	8	7	7	7
	DFKI	9	9	9	9
	UKA	10	10	10	10
	SYSTRAN	*11*	*11*	*11*	*11*
Verbatim	*Human Reference*	*1*	*1*	*1*	*1*
	UPC	2	8	4	3
	RWTH	3	6	2	5
	IBM	4	3	6	7
	IRST	4	5	4	6
	ROVER	6	4	2	2
	UW	7	2	7	4
	UKA	8	9	8	8
	SYSTRAN	*9*	*7*	*9*	*9*
ASR	RWTH	1	1	1	1
	IBM	2	2	4	3
	UPC	2	4	2	2
	IRST	4	3	2	4
	UKA	5	5	5	5
	SYSTRAN	*6*	*6*	*6*	*6*

Table 28: Ranks summary

Summary

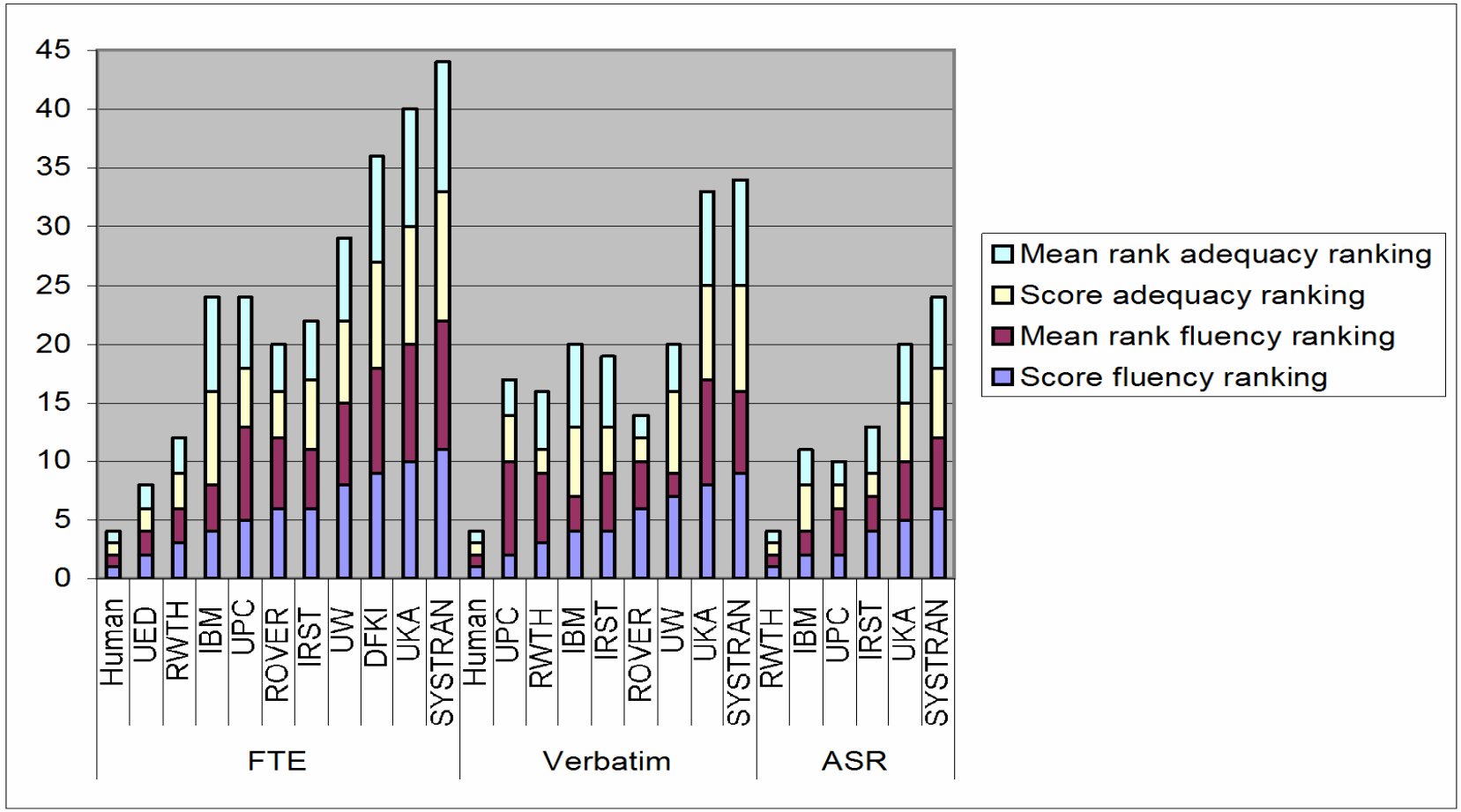


Figure 7 Ranks summary

Automatic Evaluation Metrics

- **BLEU**

  stands for BiLingual Evaluation Understudy, counts the number of word sequences (n-grams) in a sentence to be evaluated, which are common with one or more reference translations. A translation is considered better if it shares a larger number of n-grams with the reference translations. In addition, BLEU applies a penalty to those translations whose length significantly differs from that of the reference translations.

- **BLEU/NIST**

  referred to as NIST, is a variant metric of BLEU, which applies different weight for the n-grams, functions of information gain and length penalty.

- **BLEU/IBM**

  is a variant metric from IBM, with a confidence interval.

Automatic Evaluation Metrics

- **mWER**

  Multi reference Word Error Rate, computes the percentage of words which are to be inserted, deleted or substituted in the translation sentence in order to obtain the reference sentence.

- **mPER**

  Multi reference Position independent word Error Rate, is the same metric as mWER, but without taking into account the position of the words in the sentence.

- **WNM**

  The Weighted N-gram Model is a combination of BLEU and the Legitimate Translation Variation (LTV) metrics, which assign weights to words in the BLEU formulae depending on their frequency (computed using TF.IDF [9]). Only the f-measure which is a combination of the recall and the precision has been reported

- **AS-WER**

  the Word Error Rate score obtained during the alignment of the output from the ASR task with the reference translations.

Automatic results English → Spanish

- Statistics of the source documents:
  - Verbatim 28882 words 1155 sentences
  - Text 25876 words 1117 sentences
  - Asr 29531 words

Higher number of words in the manual transcription than in the FTE

Number of words in the Asr also slightly higher

Automatic results
English → Spanish

Input	Site	number of words	words per sentence	words src / words trans
ASR	IBM	31 356	27.15	0.94
	ITC-irst	30 352	26.28	0.97
	RWTH	30 643	26.53	0.96
	UKA	29 368	25.43	1.01
	UPC	29 876	25.87	0.99
	Ref-1-ver	*31 243*	*27.05*	*0.95*
Verbatim	IBM	33 134	28.69	0.87
	ITC-irst	29 022	25.13	1.00
	RWTH	29 284	25.35	0.99
	UKA	27 658	23.95	1.04
	UPC	28 661	24.81	1.01
	UW	29 170	25.26	0.99
	ROVER	*28 802*	*24.94*	*1.00*
	Ref-1-ver	*29 114*	*25.21*	*0.99*
Text	IBM	31 556	28.25	0.82
	ITC-irst	27 419	24.55	0.94
	RWTH	26 945	24.12	0.96
	UKA	26 022	23.30	0.99
	UPC	27 568	24.68	0.94
	DFKI	27 312	24.45	0.95
	UED	26 892	24.08	0.96
	UW	27 539	24.65	0.94
	ROVER	*26 285*	*23.53*	*0.98*
	Ref-1-txt	*27 032*	*24.20*	*0.96*

Table 29: LRs statistics for English-to-Spanish EPPS task

Automatic results English→Spanish

Task	Site	BLEU/ NIST	BLEU	BLEU/ IBM	mWER	mPER	WNM	AS-WER
FTE	DFKI Primary	8.70	36.32	36.33	48.06	36.36	42.66	-
	IBM Primary	9.89	47.54	47.56	41.25	31.47	48.29	-
	ITC-irst Primary	10.23	49.81	49.00	39.31	30.21	48.54	-
	ITC-irst Secondary	10.23	49.79	49.12	39.17	30.10	48.32	-
	ROVER	*10.38*	*50.74*	*49.96*	*38.15*	*29.26*	49.50	-
	RWTH Primary	10.16	49.44	49.45	39.81	30.48	48.77	-
	UED Primary	10.11	49.50	49.42	39.69	30.51	48.37	-
	UKA Primary	9.56	44.04	42.95	43.61	33.66	45.95	-
	UPC Primary	10.00	48.20	47.69	40.89	31.49	46.89	-
	UPC Secondary	10.06	48.85	48.32	40.21	31.46	47.32	-
	UW Primary	10.01	48.50	48.05	40.37	30.95	47.98	-
	SYSTRAN	*8.57*	*36.29*	*36.31*	*47.79*	*37.36*	42.10	-
Verbatim	IBM Primary	9.61	45.12	45.12	43.56	32.60	46.30	-
	ITC-irst Primary	9.91	46.61	46.33	42.19	31.51	46.34	-
	ITC-irst Secondary	9.55	44.85	44.51	44.45	33.85	46.35	-
	ROVER	*10.06*	*47.53*	*46.99*	*40.92*	*30.39*	46.84	-
	RWTH Primary	9.71	45.42	45.42	43.12	32.09	46.21	-
	UKA Primary	9.08	40.10	39.59	47.63	36.13	44.61	-
	UPC Primary	9.50	44.06	43.47	44.66	33.68	44.97	-
	UW Primary	9.24	42.57	42.52	46.15	34.84	45.38	-
	SYSTRAN	*8.10*	*32.97*	*32.97*	*51.86*	*39.74*	39.38	-
ASR	IBM Primary	8.62	35.77	35.67	52.03	38.79	43.70	51.06
	ITC-irst Primary	8.75	35.97	35.09	50.95	39.31	44.08	50.02
	ITC-irst Secondary	8.48	34.54	33.69	52.60	41.05	43.79	50.14
	RWTH Primary	8.72	35.91	35.02	50.52	38.66	43.44	50.05
	UKA Primary	8.10	31.32	30.58	55.48	43.15	41.93	56.38
	UPC Primary	8.56	34.76	34.02	51.79	40.01	43.23	50.87
	SYSTRAN	*7.03*	*23.93*	*23.86*	*62.15*	*47.84*	36.82	*61.80*

Automatic Evaluation English→Spanish

Task	Site	BLEU/ NIST	BLEU	BLEU/ IBM	mWER	mPER	WNM	AS-WER
FTE	DFKI Primary	11	12	12	12	11	12	-
	IBM Primary	9	9	9	9	8	6	-
	ITC-irst Primary	2	2	5	3	3	3	-
	ITC-irst Secondary	3	3	4	2	2	5	-
	ROVER	*1*	*1*	*1*	*1*	*1*	*1*	-
	RWTH Primary	4	5	2	5	4	2	-
	UED Primary	5	4	3	4	5	4	-
	UKA Primary	10	10	10	10	10	10	-
	UPC Primary	8	8	8	8	9	9	-
	UPC Secondary	6	6	6	6	7	8	-
	UW Primary	7	7	7	7	6	7	-
	SYSTRAN	*12*	*11*	*11*	*11*	*12*	*11*	-
Verbatim	IBM Primary	4	4	4	4	4	4	-
	ITC-irst Primary	2	2	2	2	2	3	-
	ITC-irst Secondary	5	5	5	5	6	2	-
	ROVER	*1*	*1*	*1*	*1*	*1*	*1*	-
	RWTH Primary	3	3	3	3	3	5	-
	UKA Primary	8	8	8	8	8	8	-
	UPC Primary	6	6	6	6	5	7	-
	UW Primary	7	7	7	7	7	6	-
	SYSTRAN	*9*	*9*	*9*	*9*	*9*	*9*	-
ASR	IBM Primary	3	3	1	4	2	3	3
	ITC-irst Primary	1	1	2	2	3	1	1
	ITC-irst Secondary	5	5	5	5	5	2	4
	RWTH Primary	2	2	3	1	1	4	2
	UKA Primary	6	6	6	6	6	6	6
	UPC Primary	4	4	4	3	4	5	5
	SYSTRAN	*7*	*7*	*7*	*7*	*7*	*7*	*7*

English→Spanish

- Strong correlation between the four measures
- difference between WER and PER: 9-12 %
- degradation by ASR:
  increase in PER due to word error rate of ASR
- difference between verbatim and text:
  small: BLEU=3-4%; PER=1-2%
- system combination: small improvement

Automatic Evaluation Spanish →English

Data statistics for Spanish-to-English source documents are the following:

- Text: 50 590 words, for 1782 sentences whereof
  - CORTES: 25 084 words, for 888 sentences
  - EPPS: 25 510 words, for 894 sentences
- Verbatim: 56 239 words, for 1 596 sentences whereof
  - CORTES: 28 370 words, for 699 sentences
  - EPPS: 27 873 words, for 897 sentences
- ASR: 54 708 words whereof
  - CORTES: 26 769 words.
  - EPPS: 28 939 words.

Automatic Evaluation Spanish →English

Input	Site	number of words	words per sentence	words src / words trans
ASR	IBM	62 940	39.44	0.87
	ITC-irst	61 497	38.53	0.89
	LIMSI	57 647	36.12	0.95
	RWTH	60 775	38.08	0.90
	UKA	58 840	36.87	0.93
	UPC	62 222	38.99	0.88
	Ref-1-ver	*61 207*	*38.35*	*0.89*
Verbatim	IBM	62 407	39.10	0.90
	ITC-irst	56 584	35.45	0.99
	LIMSI	55 974	35.07	1.00
	RWTH	56 168	35.19	1.00
	UKA	54 921	34.41	1.02
	UPC	57 107	35.78	0.98
	DFKI	56 802	35.59	0.99
	UW	58 065	36.38	0.97
	ROVER	*56 510*	*35.41*	*1.00*
	Ref-1-ver	*59 583*	*37.33*	*0.94*
Text	IBM	58 964	33.09	0.86
	ITC-irst	52 856	29.66	0.96
	RWTH	52 407	29.41	0.97
	UKA	50 835	28.53	1.00
	UPC	53 423	29.98	0.95
	DFKI	53 139	29.82	0.95
	UED	51 940	29.15	0.97
	UW	54 121	30.37	0.93
	ROVER	*51 486*	*28.89*	*0.98*
	Ref-1-txt	*52 051*	*29.21*	*0.97*

Table 32: LRs statistics for the Spanish-to-English task

Automatic Evaluation Spanish (Epps+Cortes) →English

Task	Site	BLEU/ NIST	BLEU	BLEU/ IBM	mWER	mPER	WNM	AS-WER
FTE	IBM Primary	10.49	48.16	48.16	41.68	30.18	44.81	-
	ITC-irst Primary	10.22	46.19	46.11	43.36	31.34	43.36	-
	ITC-irst Secondary	10.14	45.58	45.39	43.66	31.66	42.98	-
	ROVER	*10.50*	*48.07*	*48.07*	*41.64*	*30.03*	45.21	-
	RWTH Primary	10.36	47.11	47.12	42.89	30.93	44.55	-
	UED Primary	10.11	45.59	45.60	43.74	31.67	43.61	-
	UKA Primary	9.63	41.23	40.98	47.17	33.64	42.31	-
	UPC Primary	10.30	46.45	46.46	42.55	30.97	44.48	-
	DFKI Primary	9.06	37.24	37.24	63.15	34.95	39.95	-
	UW Primary	10.09	45.63	45.63	44.06	31.74	44.66	-
	SYSTRAN	*9.45*	*40.57*	*40.57*	*47.27*	*34.58*	38.39	-
Verbatim	IBM Primary	11.04	52.54	52.41	37.46	26.98	50.03	-
	ITC-irst Primary	10.57	48.85	48.49	39.94	28.66	46.06	-
	ITC-irst Secondary	10.27	47.16	46.63	41.89	30.71	45.60	-
	LIMSI Primary	9.72	42.59	42.08	44.50	31.76	41.80	-
	ROVER	11.09	52.55	52.08	36.82	26.78	50.55	-
	RWTH Primary	11.10	52.45	51.91	37.73	27.41	49.34	-
	UKA Primary	9.89	43.18	42.84	44.85	31.54	44.76	-
	UPC Primary	10.65	49.63	49.57	39.60	28.85	45.72	-
	DFKI Primary	9.09	37.50	37.42	68.61	33.53	40.34	-
	UW Primary	9.90	44.97	44.97	44.37	32.28	45.22	-
	SYSTRAN	9.89	43.73	43.74	43.54	31.19	39.66	-
ASR	IBM Primary	9.57	39.41	38.37	48.73	34.72	45.11	47.43
	ITC-irst Primary	9.03	34.30	33.92	50.50	38.58	42.96	49.04
	ITC-irst Secondary	8.64	32.52	32.24	52.52	40.82	42.46	49.43
	LIMSI Primary	8.48	32.60	32.13	52.58	38.18	38.28	51.72
	RWTH Primary	9.26	36.13	35.79	49.00	38.02	43.96	47.66
	UKA Primary	8.42	30.13	29.79	54.82	41.42	40.96	54.63
	UPC Primary	9.04	34.83	34.30	51.01	39.01	40.87	49.57
	SYSTRAN	8.06	29.22	29.22	62.27	47.77	36.60	*62.14*

Table 33: Evaluation results for the Spanish-to-English task

Automatic Evaluation

Spanish →English

Task	Site	BLEU/ NIST	BLEU	BLEU/ IBM	mWER	mPER	WNM	AS-WER
FTE	IBM Primary	2	1	1	2	2	2	-
	ITC-irst Primary	5	5	5	5	5	7	-
	ITC-irst Secondary	6	8	8	6	6	8	-
	ROVER	*1*	*2*	*2*	*1*	*1*	1	-
	RWTH Primary	3	3	3	4	3	4	-
	UED Primary	7	7	7	7	7	6	-
	UKA Primary	9	9	9	9	9	9	-
	UPC Primary	4	4	4	3	4	5	-
	DFKI Primary	11	11	11	11	11	10	-
	UW Primary	8	6	6	8	8	3	-
	SYSTRAN	*10*	*10*	*10*	*10*	*10*	11	-
Verbatim	IBM Primary	3	2	1	2	2	2	-
	ITC-irst Primary	5	5	5	5	4	4	-
	ITC-irst Secondary	6	6	6	6	6	6	-
	LIMSI Primary	10	10	10	9	9	9	-
	ROVER	2	1	2	1	1	1	-
	RWTH Primary	1	3	3	3	3	3	-
	UKA Primary	8	9	9	10	8	8	-
	UPC Primary	4	4	4	4	5	5	-
	DFKI Primary	11	11	11	11	11	10	-
	UW Primary	7	7	7	8	10	7	-
	SYSTRAN	9	8	8	7	7	11	-
ASR	IBM Primary	1	1	1	1	1	1	1
	ITC-irst [illegible]	4	4	4	3	4	3	4
	ITC-irst Secondary	5	6	5	5	6	4	5
	LIMSI Primary	6	5	6	6	3	7	6
	RWTH Primary	2	2	2	2	2	2	2
	UKA Primary	7	7	7	7	7	5	7
	UPC Primary	3	3	3	4	5	6	3
	SYSTRAN	*8*	*8*	*8*	*8*	*8*	8	*8*

Table 34: Ranking of systems for the Spanish-to-English task

Automatic Evaluation
Spanish →English

- Spanish Cortes and EPPS have been evaluated separately:
  - Results on EPPS are better than those from the Cortes
  - The ranking does not vary
- comparison with English→ Spanish: better by about 6% (again!)
- ASR condition: IBM better by 3% in BLEU and PER
- difference between WER and PER: 13%
  (? exception: DFKI with 26-30%)
- degradation by ASR:
- increase in PER: = WER of Asr
- difference between verbatim and text: small
- system combination: virtually no improvement

Automatic Evaluation

Spanish (Cortes)→English

- Difference to EPPS: worse by 5%
- ASR condition: IBM better by 3% in BLEU and PER
- Difference between WER and PER: 14-16 %

  (? exception DFKI: verbatim = 40%, whereas text = 13%)
- degradation by ASR:

  increase in PER: = WER of Asr
- difference between verbatim and text:

  about 3% (BLEU,PER,WER)
- System combination: no improvement

Automatic Evaluation Chinese →English

- Data statistics for Chinese→English sources:
  - Verbatim 27730 words for 1232 sentences

Input	Site	number of words	words per sentence	words src / words trans
ASR	ITC-irst	30 584	24.82	0.89
	RWTH	30 198	24.51	0.91
	UKA	31 815	25.82	0.86
	ICT	29 618	24.04	0.92
	NLPR	32 216	26.15	0.85
	Ref-1-ver	31 184	25.31	0.88
Verbatim	ITC-irst	28 648	23.25	0.96
	RWTH	28 541	23.17	0.96
	UKA	27 996	22.72	0.98
	ICT	27 666	22.46	0.99
	NLPR	32 283	26.20	0.85
	NRC	29 971	24.33	0.91
	Ref-1-ver	30 707	24.92	0.89

Table 39: LRs statistics for the Chinese-to-English VOA task

Automatic Evaluation Chinese→English

Task	Site	BLEU/ NIST	BLEU	BLEU/ IBM	mWER	mPER	WNM	AS-WER
Verbatim	ICT Primary	6.03	13.70	13.20	78.68	58.06	25.72	-
	ITC-irst Primary	6.01	14.04	13.49	79.76	59.42	25.73	-
	ITC-irst Secondary	6.00	13.92	13.42	80.05	59.66	25.70	-
	NLPR Secondary	4.35	7.30	7.30	102.92	79.82	22.25	-
	NRC Primary	5.49	12.24	12.25	84.50	63.07	26.43	-
	NRC Secondary	5.80	12.76	12.76	84.23	61.67	27.36	-
	RWTH Primary	6.45	16.07	15.32	78.08	56.34	27.58	-
	UKA Primary	5.51	10.81	10.30	82.22	61.72	24.21	-
	SYSTRAN	*4.28*	*6.53*	*6.53*	*95.37*	*74.77*	23.35	-
ASR	ICT Primary	4.90	10.86	10.46	77.79	62.46	24.56	83.31
	ITC-irst Primary	4.92	11.07	10.83	78.80	63.19	24.71	83.58
	ITC-irst Secondary	4.95	11.12	10.88	78.93	63.22	24.69	83.78
	NLPR Secondary	4.09	6.74	6.74	87.28	71.27	20.85	90.49
	RWTH Primary	5.17	12.39	12.09	77.99	61.98	26.47	83.02
	UKA Primary	4.59	8.46	8.46	82.92	66.89	23.86	88.28
	SYSTRAN	4.38	8.62	8.48	80.59	65.78	22.52	95.20

Automatic Evaluation Chinese→English

Task	Site	BLEU/ NIST	BLEU	BLEU/ IBM	mWER	mPER	WNM	AS-WER
Verbatim	ICT Primary	6.03	13.70	13.20	78.68	58.06	25.72	-
	ITC-irst Primary	6.01	14.04	13.49	79.76	59.42	25.73	-
	ITC-irst Secondary	6.00	13.92	13.42	80.05	59.66	25.70	-
	NLPR Secondary	4.35	7.30	7.30	102.92	79.82	22.25	-
	NRC Primary	5.49	12.24	12.25	84.50	63.07	26.43	-
	NRC Secondary	5.80	12.76	12.76	84.23	61.67	27.36	-
	RWTH Primary	6.45	16.07	15.32	78.08	56.34	27.58	-
	UKA Primary	5.51	10.81	10.30	82.22	61.72	24.21	-
	SYSTRAN	*4.28*	*6.53*	*6.53*	*95.37*	*74.77*	23.35	-
ASR	ICT Primary	4.90	10.86	10.46	77.79	62.46	24.56	83.31
	ITC-irst Primary	4.92	11.07	10.83	78.80	63.19	24.71	83.58
	ITC-irst Secondary	4.95	11.12	10.88	78.93	63.22	24.69	83.78
	NLPR Secondary	4.09	6.74	6.74	87.28	71.27	20.85	90.49
	RWTH Primary	5.17	12.39	12.09	77.99	61.98	26.47	83.02
	UKA Primary	4.59	8.46	8.46	82.92	66.89	23.86	88.28
	SYSTRAN	4.38	8.62	8.48	80.59	65.78	22.52	95.20

Automatic Evaluation Chinese→English

Task	Site	BLEU/ NIST	BLEU	BLEU/ IBM	mWER	mPER	WNM	AS-WER
Verbatim	ICT Primary	2	4	4	2	2	5	-
	ITC-irst Primary	3	2	2	3	3	4	-
	ITC-irst Secondary	4	3	3	4	4	6	-
	NLPR Secondary	9	9	9	9	9	9	-
	NRC Primary	7	6	6	8	8	3	-
	NRC Secondary	5	5	5	7	5	2	-
	RWTH Primary	1	1	1	1	1	1	-
	UKA Primary	6	8	8	6	6	7	-
	SYSTRAN	*8*	*7*	*7*	*5*	*7*	8	-
ASR	ICT Primary	4	4	4	1	2	4	2
	ITC-irst Primary	3	3	3	3	3	2	3
	ITC-irst Secondary	2	2	2	4	4	3	4
	NLPR Secondary	7	7	7	7	7	7	6
	RWTH Primary	1	1	1	2	1	1	1
	UKA Primary	5	5	5	6	6	5	5
	SYSTRAN	*6*	*6*	*6*	*5*	*5*	6	7

Automatic Evaluation Chinese →English

- Absolute performance: much worse than Spanish (in both directions)

  (BLEU: 12-15%; PER: 56-64%)
- difference between WER and PER: 17-20%
- degradation by ASR:

  increase in PER: = less than CER of Asr

Data Analysis

Metric	En->Es			Es->En			Zh->En	
	ASR	Text	Verb	ASR	Text	Verb	ASR	Verb
BLEU ↔ IBM	99.75	99.74	99.85	99.86	99.97	99.90	99.91	99.74
BLEU ↔ mPER	98.74	98.76	97.87	86.90	98.23	96.11	95.94	93.96
BLEU ↔ WNM	98.58	97.79	96.61	80.68	88.49	90.30	94.95	87.46
IBM ↔ mPER	99.16	98.65	97.91	85.54	98.24	95.89	95.34	93.21
IBM ↔ WNM	97.85	97.93	96.52	81.06	88.32	89.72	95.68	89.90
mPER ↔ WNM	94.79	98.64	91.35	76.03	92.68	86.43	91.12	84.25

Table 42: Pearson correlation between metrics scoring

- All the metrics are strongly correlated
- Bleu and Bleu/IBM scores almost the same

Automatic metrics and human evaluation

- Automatic metrics compared with human evaluation results
- English→Spanish direction
- Correlations between automatic metrics' scores and fluency/adequacy scores
- Hamming distance between automatic metrics' ranks and fluency/adequacy ranks

Automatic metrics and human evaluation

Metrics	ASR scoring	Text scoring	Verb scoring	ASR ranking	Text ranking	Verb ranking
BLEU vs Fluency	97.91	77.16	95.15	4	10	5
IBM vs Fluency	97.26	78.35	94.70	4	9	5
mPER vs Fluency	96.04	80.85	90.25	2	9	5
WNM vs Fluency	98.66	78.82	95.69	3	9	5
BLEU vs Adequacy	97.09	80.00	95.53	4	9	4
IBM vs Adequacy	95.80	80.38	94.95	3	8	4
mPER vs Adequacy	94.47	83.73	90.31	3	9	4
WNM vs Adequacy	98.45	80.67	97.49	4	8	5

Best Results

Task	Condition	BLEU [%]	NIST	PER [%]	WER [%]
E → S: EPPS	ASR (WER=6.9%)	36.0	8.75	39.3	51.0
	Verbatim	46.6	9.91	31.5	42.2
	Text	49.8	10.23	30.2	39.3
S → E: EPPS	ASR (WER=6.2%)	42.8	9.65	32.6	45.9
	Verbatim	55.2	10.91	25.7	36.3
	Text	54.1	10.77	26.4	36.2
S → E: CORTES	ASR (WER=9.8%)	33.6	8.37	39.1	56.2
	Verbatim	46.2	9.85	31.7	44.9
	Text	42.1	9.26	34.1	47.3
C → E: BN	ASR (CER=9.8%)	12.4	5.17	62.0	78.0
	Verb.=Text	16.3	6.43	56.0	77.6

General observations

- Strong correlation between all automatic measures
- Comparison of Tasks
  - best task: S→E EPPS
  - S→E CORTES worse: -11% BLEU, +6% PER
  - E→S EPPS: similar
- Verbatim versus text comparison:
  - virtually no difference
  - verbatim sometimes is better !
- Asr versus verbatim comparison:
  - degradation in PER: = WER of Asr
  - degradation in BLEU: slightly more
- Chinese: worse performance
  - (bigger) mismatch between training and test
  - different language structures!

TTS Evaluation

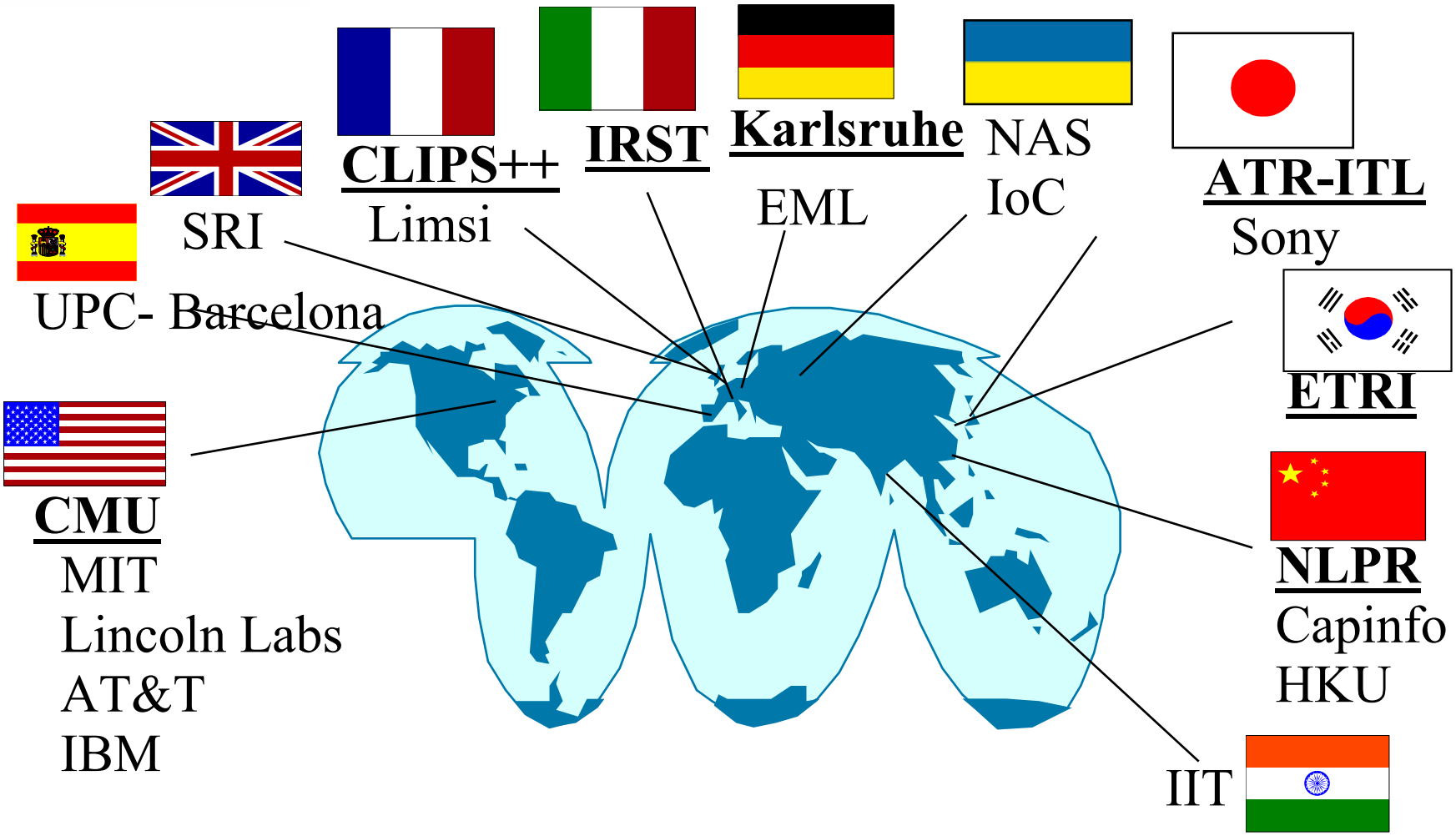
- Partnership with ECESS
- Less formalized framework compared to ASR and SLT
- Tasks aims differs:
  - to evaluate globally TTS systems
  - to analyze components (diagnostic tests)

TTS Evaluation

Evaluation task		Languages
Text processing		***CN, EN***
M1.1	Non Standard Word Normalization	EN
M1.2	End-of-sentence detection (EN) Words segmentation (CN)	 CN, EN
M1.3	POS (Part-Of-Speech) Tagging	CN, EN
M1.4	Grapheme-to-phoneme conversion	CN, EN
Prosody Generation		***CN, EN, ES***
M2.1	Evaluation of prosody – Use of segmental information	EN, ES
M2.2	Evaluation of prosody – Rating of delexicalised utterances	CN, EN, ES
M2.3	Evaluation of prosody – Choice of a delexicalised utterance	CN, EN, ES
Acoustic Synthesis		***CN, EN***
M3.1	Intelligibility test (Semantically Unpredictable Sentences)	CN, EN
M3.2	Judgment test (Intelligibility and Naturalness)	CN, EN
Intra-lingual Voice Conversion (IVC)		***CN, EN, ES***
VC1	Comparison of speaker identities	
VC2	Evaluation of overall speech quality	
Crosslingual Voice Conversion (CVC)		***EN/ES, ES/EN***
VC1	Comparison of speaker identities	
VC2	Evaluation of overall speech quality	

Thank you !

C-STAR, now


SRI
CLIPS++
Limsi
IRST
Karlsruhe
EML
NAS
IoC
ATR-ITL
Sony
UPC- Barcelona
ETRI
CMU
MIT
Lincoln Labs
AT&T
IBM
NLPR
Capinfo
HKU
IIT