MT EVALUATION & TC-STAR

Khalid CHOUKRI, Olivier Hamon, Djamel Mostefa

ELRA/ELDA
55 Rue Brillat-Savarin, F-75013 Paris, France
Tel. +33 1 43 13 33 33 -- Fax. +33 1 43 13 33 30

Email: choukri@elda.org **Web** http://www.elda.org/ **or** http://www.elra.info/

Presentation Outline Key themes

- Part 1: general views on Evaluation(s)
- Part 2: TC-STAR evaluations approaches & lessons
- Part 3 (1slide): some (of my) open issues for

TC-STAR evaluations
……. 3 Consecutives annual evaluations

1.SLT in the following directions

i. Chinese-to-English (Broadcast News)

ii. Spanish-to-English (European Parliament plenary speeches)

iii. English-to-Spanish (European Parliament plenary speeches)

2.ASR in the following languages

i. English (European Parliament plenary speeches)

ii. Spanish (European Parliament plenary speeches)

iii. Mandarin Chinese (Broadcast News)

3.TTS in Chinese, English, and Spanish under the following conditions:

i. Complete system

ii. Voice conversion intralingual and crosslingual, expressive speech:

iii. Component evaluation

SLT Tasks

- 3 Inputs
  - **ASR**: translate automatic transcripts from ASR engines (ROVERed), with case and punctuation, **no manual segmentation**
  - **Verbatim**: translate manual transcripts, with case and punctuation
  - **Text**: translate Final Text Edition (FTE) documents, with case and punctuation
- 2 Conditions
  - **Primary**: use single-best hypo from ASR output, use only for training:
    - EPPS: EPPS training set
    - CORTES: Spanish Parliament training set
    - VOA: LDC Large Data
  - **Secondary**: like primary plus ASR word graphs or any other optional input and publicly available data, and use any publicly available data for training

SLT inputs example

Verbatim

I'm I'm I'm starting to know what Frank Sinatra must have felt like .

ASR output

and I'm times and starting to know what Frank Sinatra must have felt like .

Text

I am starting to know what Frank Sinatra must have felt like,

Resources Production

Development data:

- *2005 development data*
- *2005 test data*
- *2006 development data*
- *2006 test data*

Test	
EPPS	manual transcripts (ELDA) taken from ASR test ~25000 words + reference translations (ELDA) 200 kwords (En →Es + Es→En)
CORTES	manual transcripts (ELDA) taken from ASR test ~25000 words + reference translations (ELDA) 100 kwords (Es→En)
VOA	manual transcripts (ELDA) of 3h excerpt from ASR test ~25000 words + English reference translations (ELDA) 50 kwords (Zh→En)

Validation procedure for SLT data sets

- **Task: assess quality of given translations (references),**
- **Per set: 1200 contingent words (5%) are selected (from different texts, from source text,except Mandarin; there from target text)**
- **Two translations per text from different agencies, same samples, checked by a professional translator**
- **Procedure and criteria are adopted from LDC/NIST**
- **Max. 40 penalty points per translation allowed**
- **Validators were unaware of the scoring,**

  only of categories
-

Error	Penalty
Syntactical	3
Lexical	3
Poor usage	1
Capitalisation	1
Punctuation - spelling errors	0.5 (max 10)

Validation results for SLT data sets (should be < 40)

Direction	Data	Task	Agency 1	Agency 2
English-to-Spanish	EPPS	FTE	35	14
	EPPS	Verbatim	40	17
Spanish-to-English	EPPS	FTE	18 R	38 R
	EPPS	Verbatim	20	40
	PARL	FTE	34 R	35 R
	PARL	Verbatim	26.5 R	22.5 R
Chinese-to-English	VOA	Verbatim	27 R	37 R

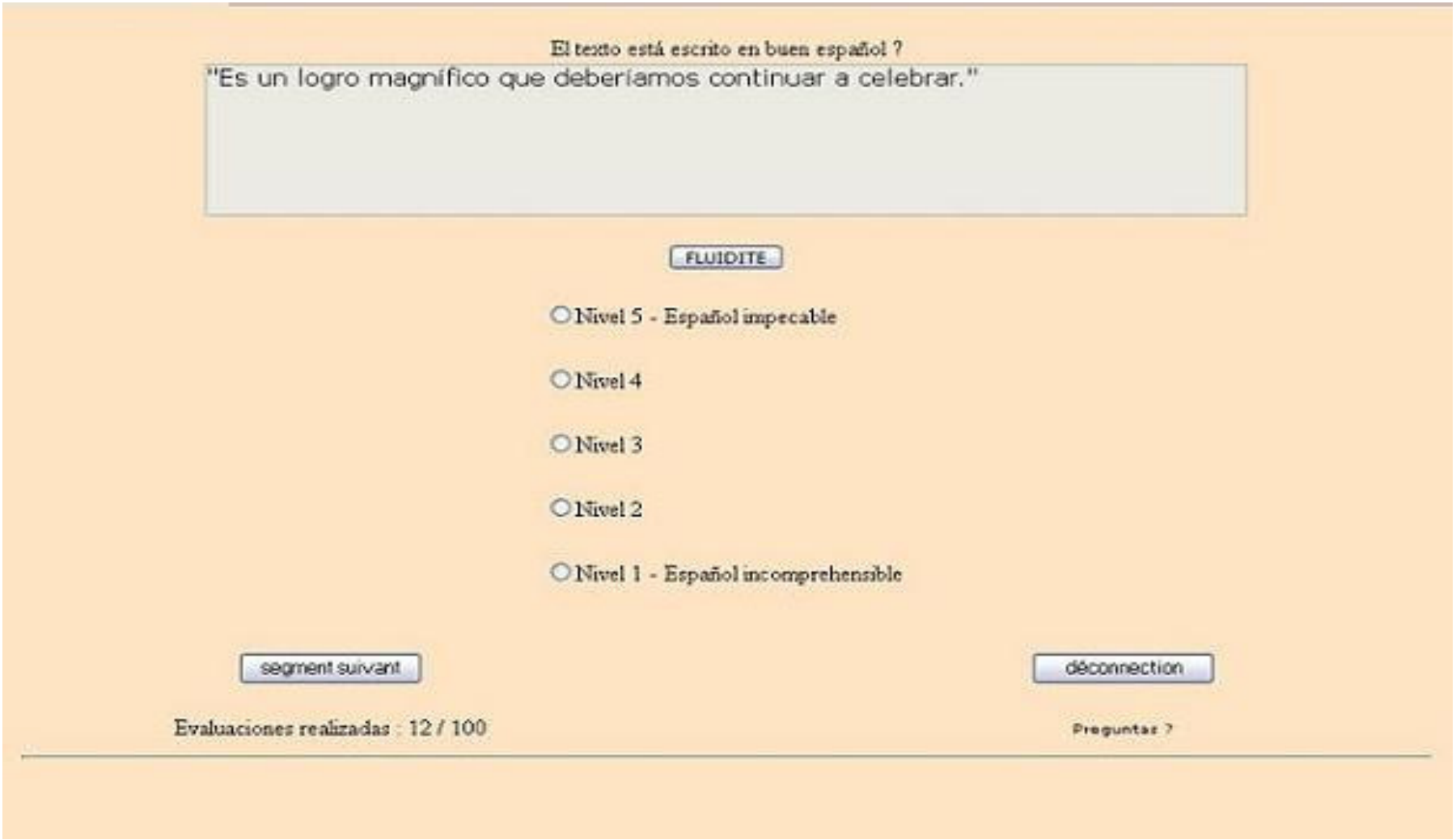
SLT Scoring

- Automatic metrics:
  - BLEU, NIST, IBM(BLEU), mWER, mPER, WNM
- Human evaluation
  - English-to-Spanish direction
  - 100 evaluators (Spanish native speakers, university level education)
  - Around 350 segments by primary system + Softissimo + Systran + one reference translation (45 systems)
  - Evaluation of adequacy and fluency
    - Evaluation of adequacy: the target segment is compared to a reference segment
    - Evaluation of fluency: only the quality of “grammar” is evaluated.
  - Each segment assessed twice by two different judges

SLT Scoring

- Human evaluation
  - 15 835 segments are evaluated, which correspond to 317 segments per evaluator.
  - Re-use of a specific web interface which has already been used for the human evaluation of the French CESTA project.
  - The evaluation is done online

Human Evaluation Interface



SLT Participants

Number of submissions all types of input and all training conditions included:

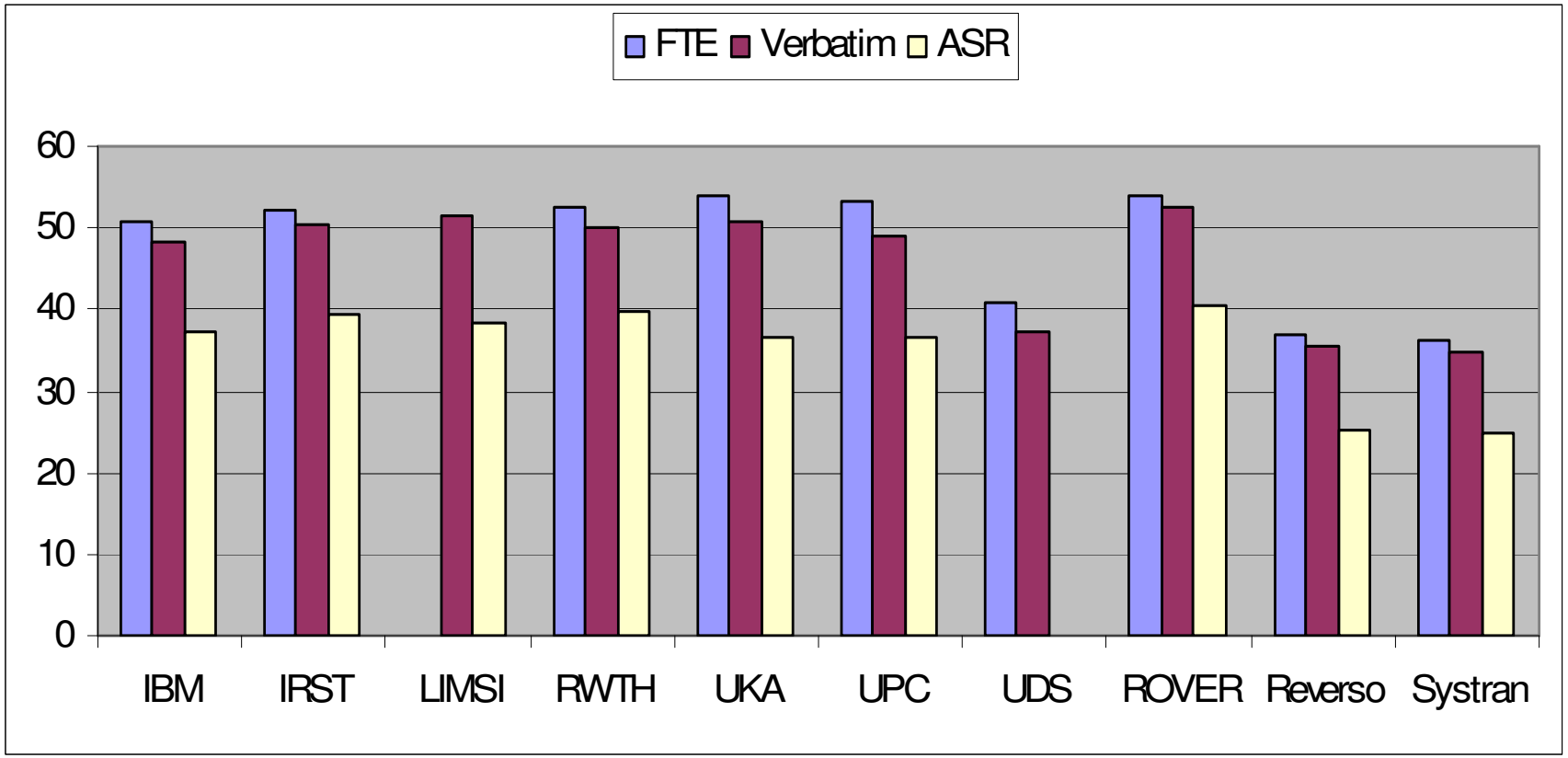
- ***Final Text Edition, Verbatim, ASR***
- ***(primary, secondary)***
- ***One combined TC-STAR « system » and Softissimo and Systran***

Languages	2007	2006	2005
En->Es	57	40	28
Es->En	64	48	38
Zh->En	55	33	31
Total	176	121	97

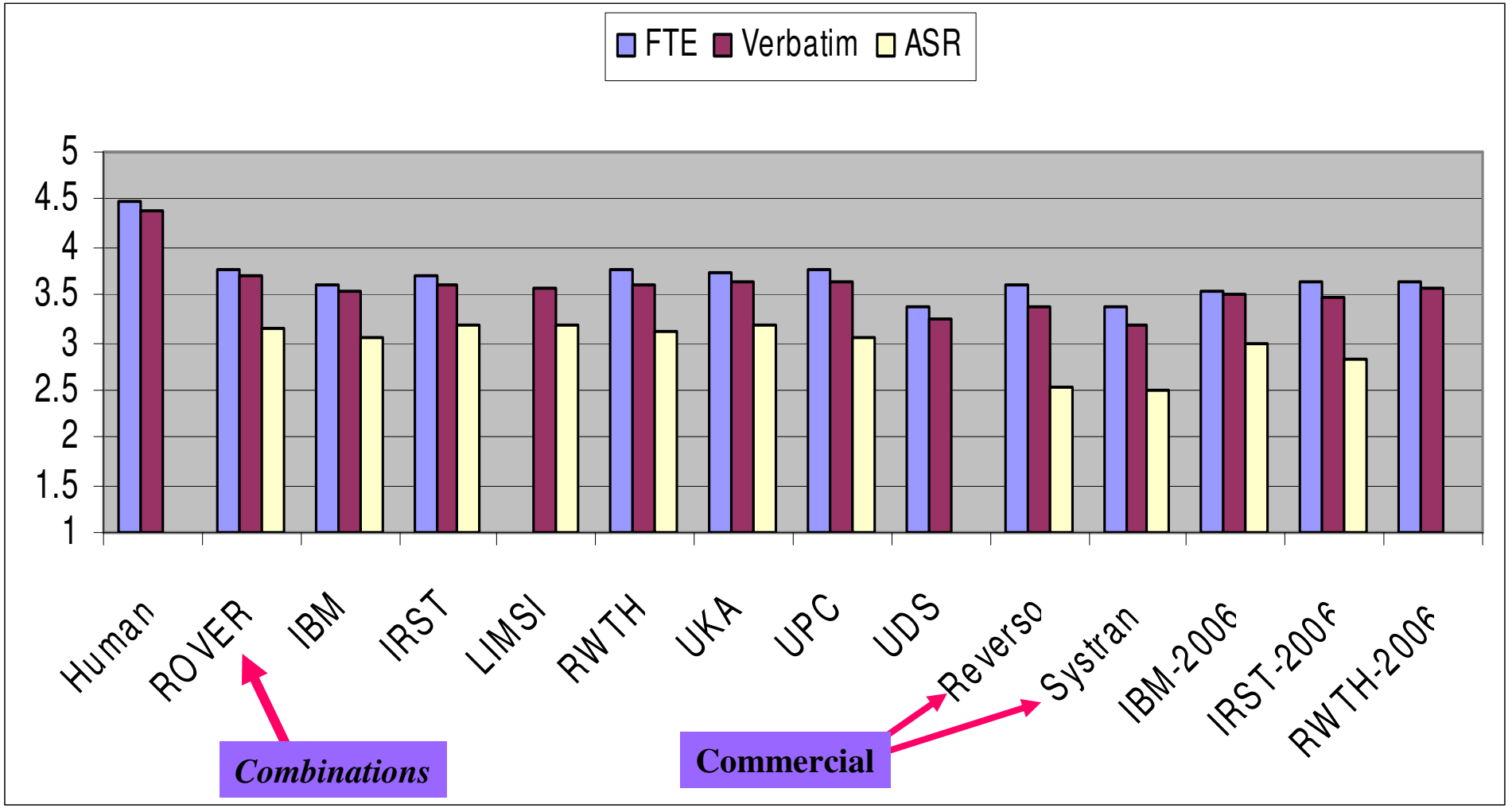
SLT Results … Summary

- See details in the MT paper by O. Hamon et. al. (tomorrow)
- Here just a summary to connect with Human evaluation(s)
- Use of **Systran Premium 5.0 Global Pack (800€),**
    - Acquired via Internet
    - No customization

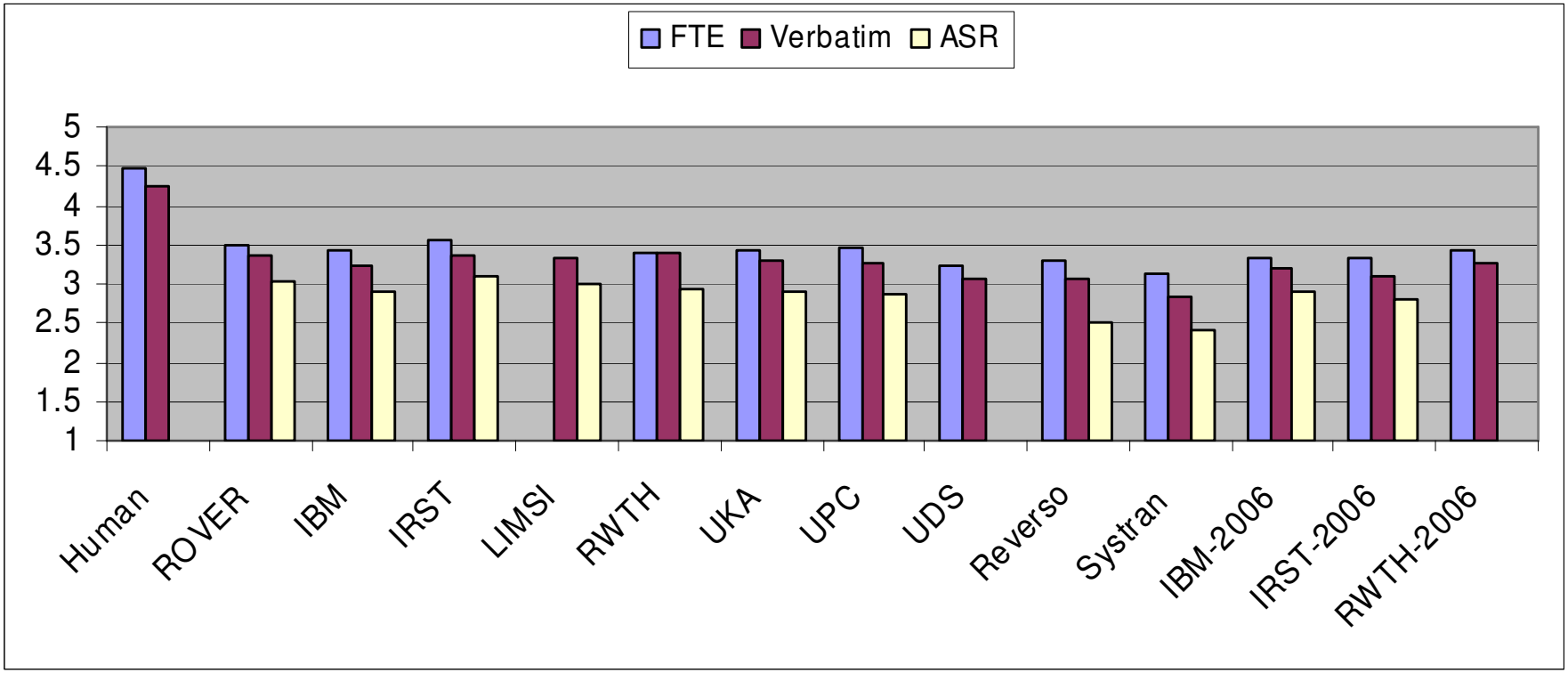
BLEU/NIST Results – EnEs (scale: 0-100)



Human Evaluation Translations … EnEs adequacy (1-5)



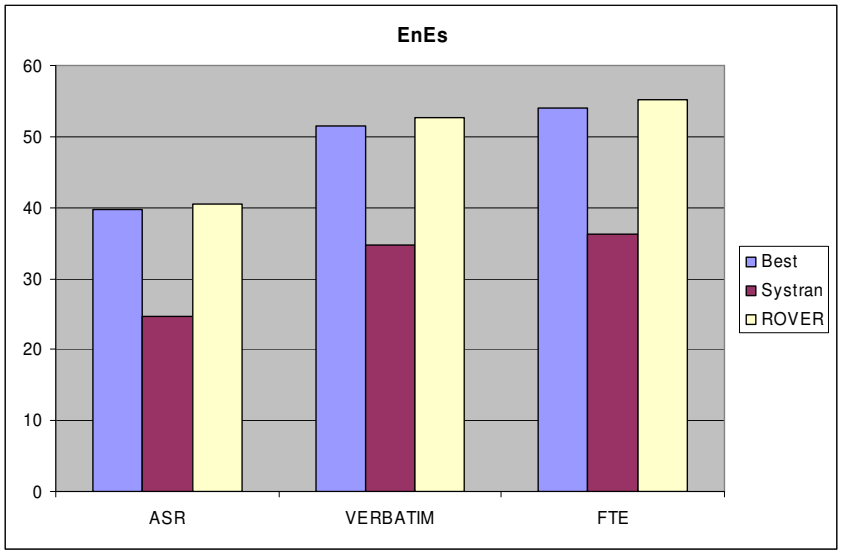
Human Eval Results fluency (0-100)

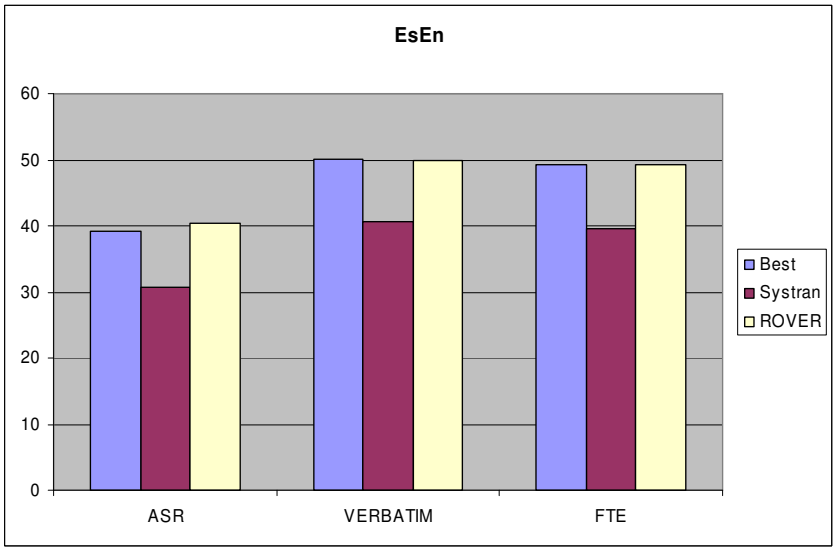


Human Eval Results – subset correlation

	NIST	BLEU	IBM	mWER	mPER	WNM
FTE	99.88	99.77	99.76	99.89	99.93	99.72
Verbatim	99.92	99.89	99.84	99.82	99.90	99.84
ASR	99.96	99.91	99.93	99.94	99.93	99.42

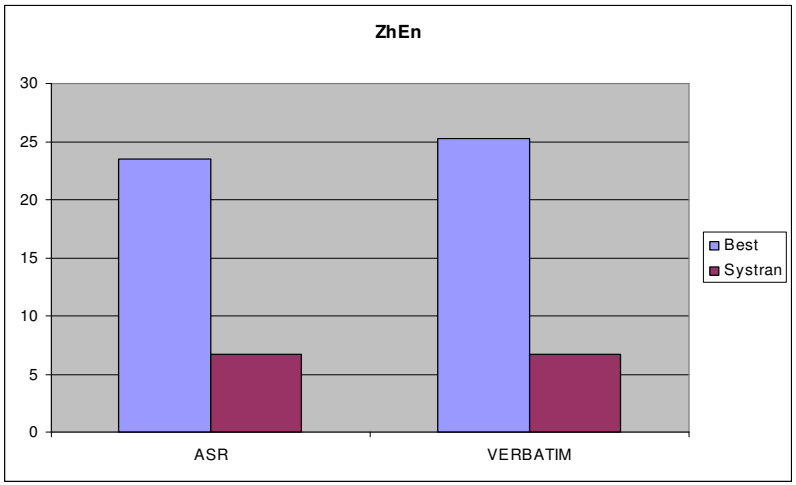
SLT Automatic Results … BLEU measures… TC-STAR versus ….





EPPS

Broadcast News



Automatic Metrics Comparison

Metric	En->Es			Es->En			Zh->En	
	ASR	Text	Verb	ASR	Text	Verb	ASR	Verb
BLEU ↔ IBM	99.94	99.90	99.90	99.74	99.96	99.94	99.99	99.99
BLEU ↔ mPER	97.93	97.72	98.19	94.30	88.77	94.10	97.37	96.09
BLEU ↔ WNM	99.25	98.93	97.01	96.33	96.18	95.03	96.75	97.81
IBM ↔ mPER	97.56	97.50	98.06	93.34	87.88	93.96	97.28	95.95
IBM ↔ WNM	99.55	99.24	96.87	95.96	95.88	95.22	96.58	98.02
mPER ↔ WNM	96.57	96.56	94.41	84.42	81.87	83.49	92.10	91.29

up: scoring correlation ; down: ranking correlation

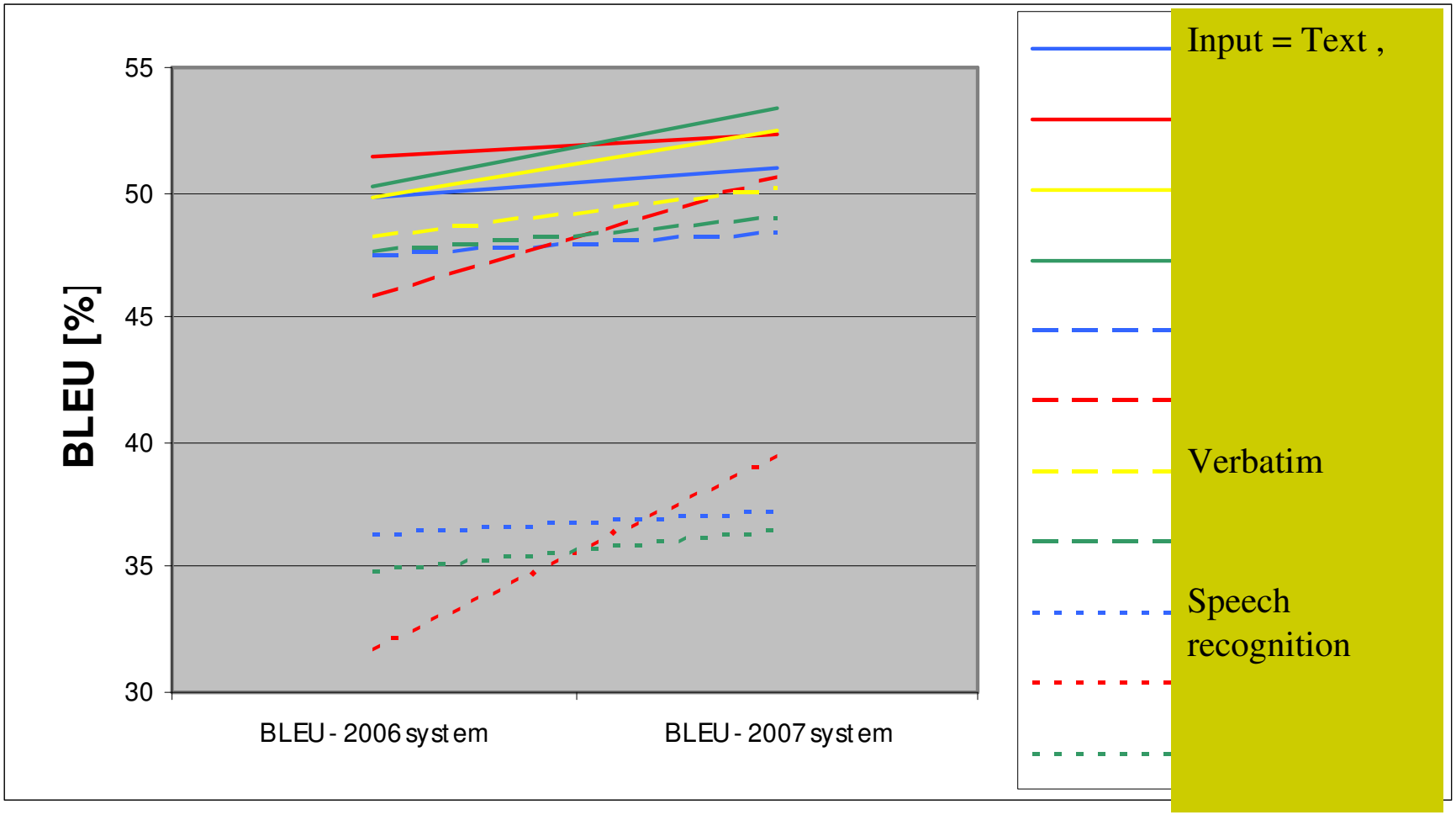
Correlations of automatic Metrics vs Human Evaluation (EnEs)

Metrics	ASR	Text	Verb
BLEU vs. Fluency	98.16	86.68	92.93
IBM vs. Fluency	98.47	86.71	92.51
mPER vs. Fluency	94.87	78.1	85.62
WNM vs. Fluency	98.97	87.85	94.34
BLEU vs. Adequacy	97.26	84.23	93.83
IBM vs. Adequacy	97.46	84.13	93.46
mPER vs. Adequacy	96.57	73.74	87.14
WNM vs. Adequacy	98.48	81.19	89.36

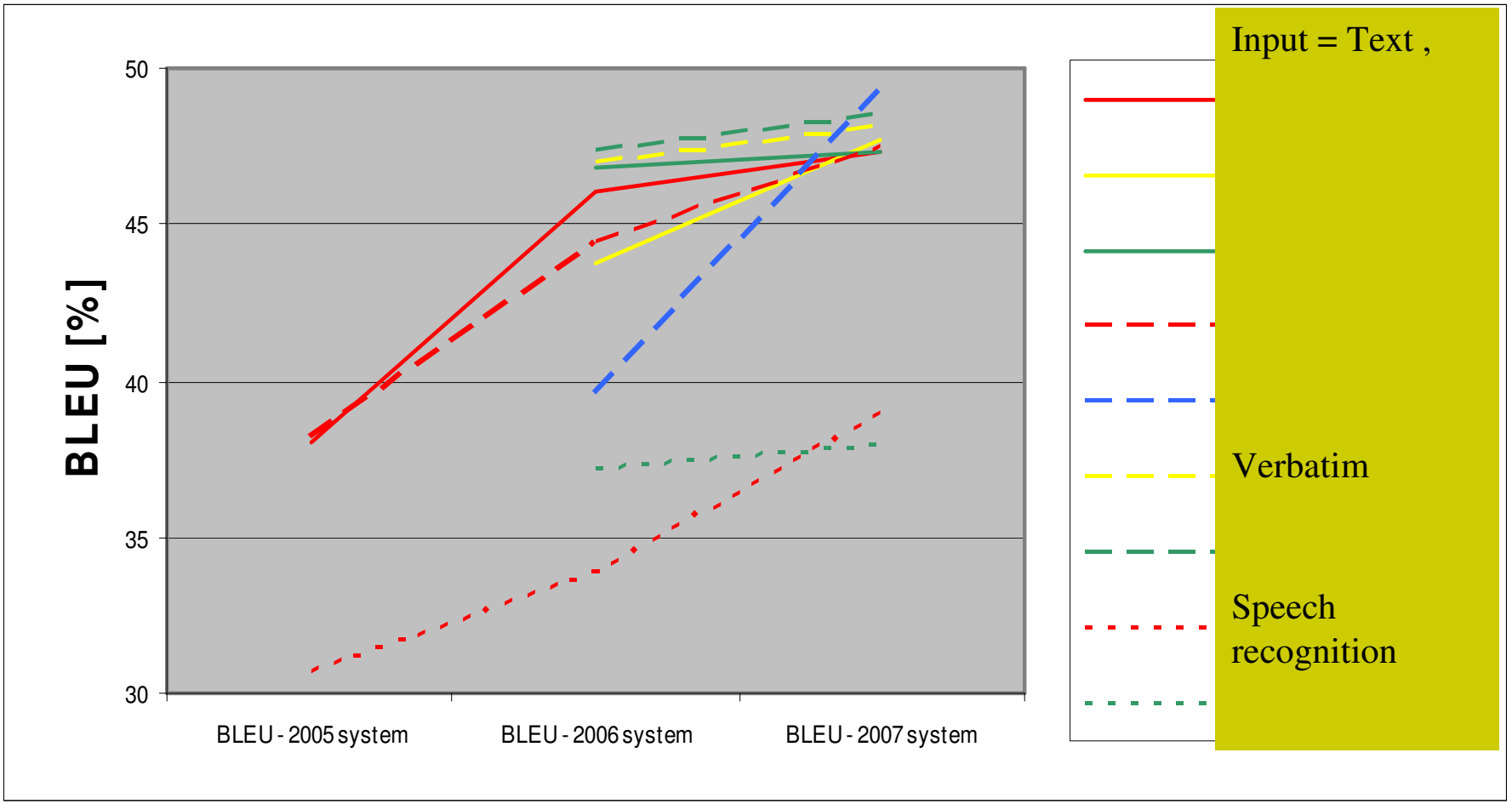
up: scoring correlation ; down: ranking correlation

- What are our conclusions?
- For whom?

Improvement of SLT Performances (En→Es)



Improvement of SLT Performances (Es→En)



End-to-End evaluation

End-to-End

- The end-to-end evaluation is carried out for 1 translation direction: English-to-Spanish
- Evaluation of ASR (Rover) + SLT (Rover) +TTS (UPC) system
- Same segments as for SLT human evaluation
- Evaluation tasks:
  - **Adequacy**: comprehension test ➔ *Very surprising results!*
  - **Fluency**: judgement test with several questions related to fluency and also usability of the system

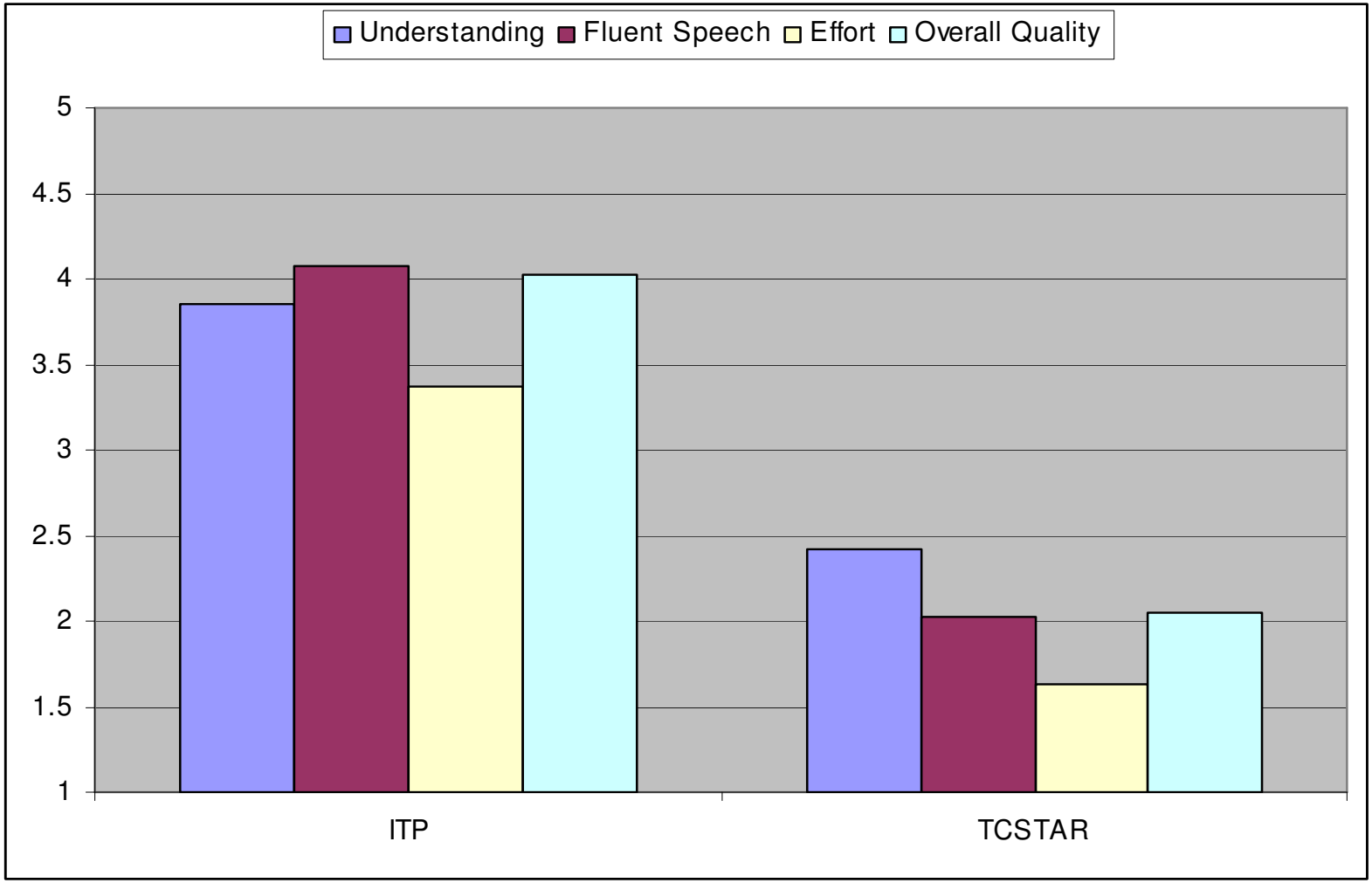
End-to-End (2/2)

- Test data:
  - Input: audio data: 20 * 3 minutes of speech in English. For each segment:
    - The ASR ROVER output (English).
    - The ASR ROVER output (Spanish).
    - The synthesis (TTS) by UPC (Spanish).
  - The speech from the interpreter (ITP) is collected and evaluated as a “top-line”.

Fluency questionnaire

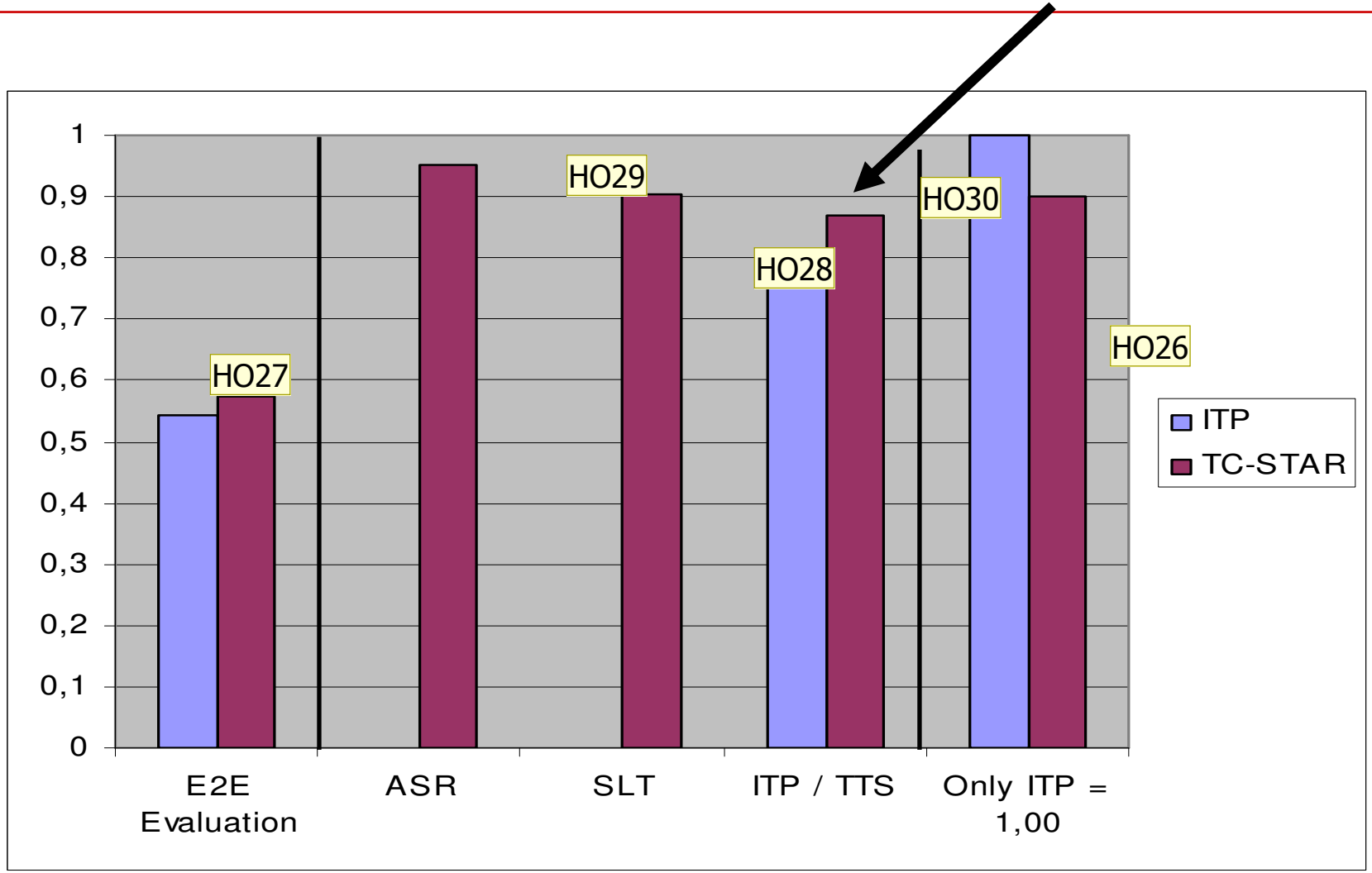
- [Understanding] Do you think that you have understood the message?
  1: Not at all , ...5: Yes, absolutely
- [Fluent Speech] Is the speech in good Spanish?
  1: No, it is very bad ..... 5: Yes, it is perfect
- [Effort] Rate the listening effort
  1: Very high ..... 5: Low, as natural speech
- [Overall Quality] Rate the overall quality of this audio sample
  1: Very badm unusable ..... 5: It is very useful

End to End results (subjective test: 1…5)



End to End results from 2006 tests
The Key Lesson from TC-STAR

Comprehension test: 0-1)



Folie 29

HO26 -2 evaluated systems: ITP for the interpreter version and TC-STAR for the automatic speech-to-speech translation system
Hamon Olivier; 16.06.2006

HO27 - E2E Evaluation: the evaluation was done by the same assessors who did the subjective evaluation.
Hamon Olivier; 16.06.2006

HO28 - ITP / TTS: as it was not foreseen that results would be better for TC-STAR than for ITP, the audio files had been validated to check whether they contained the answers to the questions. The first conclusions that can be drawn from this are: it was difficult for the assessors to find the answers (questions too hard?) and as the interpreter selects and reformulates the information, missing some details, then the question becomes too specific and not appropriate.
Hamon Olivier; 16.06.2006

HO29 - TTS, SLT, ASR: in order to determine where the information was lost for the TC-STAR system, files from each component (recognized files for ASR, translated files for SLT, synthethized files for TTS) have been checked. The overall loss is 15% of the information, 5% being lost at each step.
Hamon Olivier; 16.06.2006

HO30 - Only ITP: in the end, we used the questions whose answers were included in the interpreter files. So the TC-STAR system lost 10% of the information regarding the ITP evaluation (instead of 15%).
Hamon Olivier; 16.06.2006

TC-STAR Tasks

- **More results from the 2007 Campaign**

  http://www.tc-star.org/

- **Evaluation packages available ☺**

  E0002 TC-STAR Evaluation Package - ASR English

  E0003 TC-STAR Evaluation Package - ASR Spanish

  E0004 TC-STAR Evaluation Package - ASR Mandarin Chinese

  E0005 TC-STAR Evaluation Package - SLT English-to-Spanish

  E0006 TC-STAR Evaluation Package - SLT Spanish-to-English

  E0007 TC-STAR Evaluation Package - SLT Chinese-to-English

Some open issues/topics

1. "Role of the user in the evaluation process of MT (& Assisted MT); How much difficult taking the user in the evaluation process?"
2. "How to measure MT performance/success: user satisfaction or technology accuracy?" – do they correlate ?
3. "How to quantify the success in each situation?
4. How much is it dependent from scenarios and context (application)?"
5. "Are the best (performance) systems the most successful commercially?" –
6. "How useful the evaluation is? Pushing a head the knowledge or killing the innovation?"
7. ……

Thank you very much for you attention

Slides will be made available