



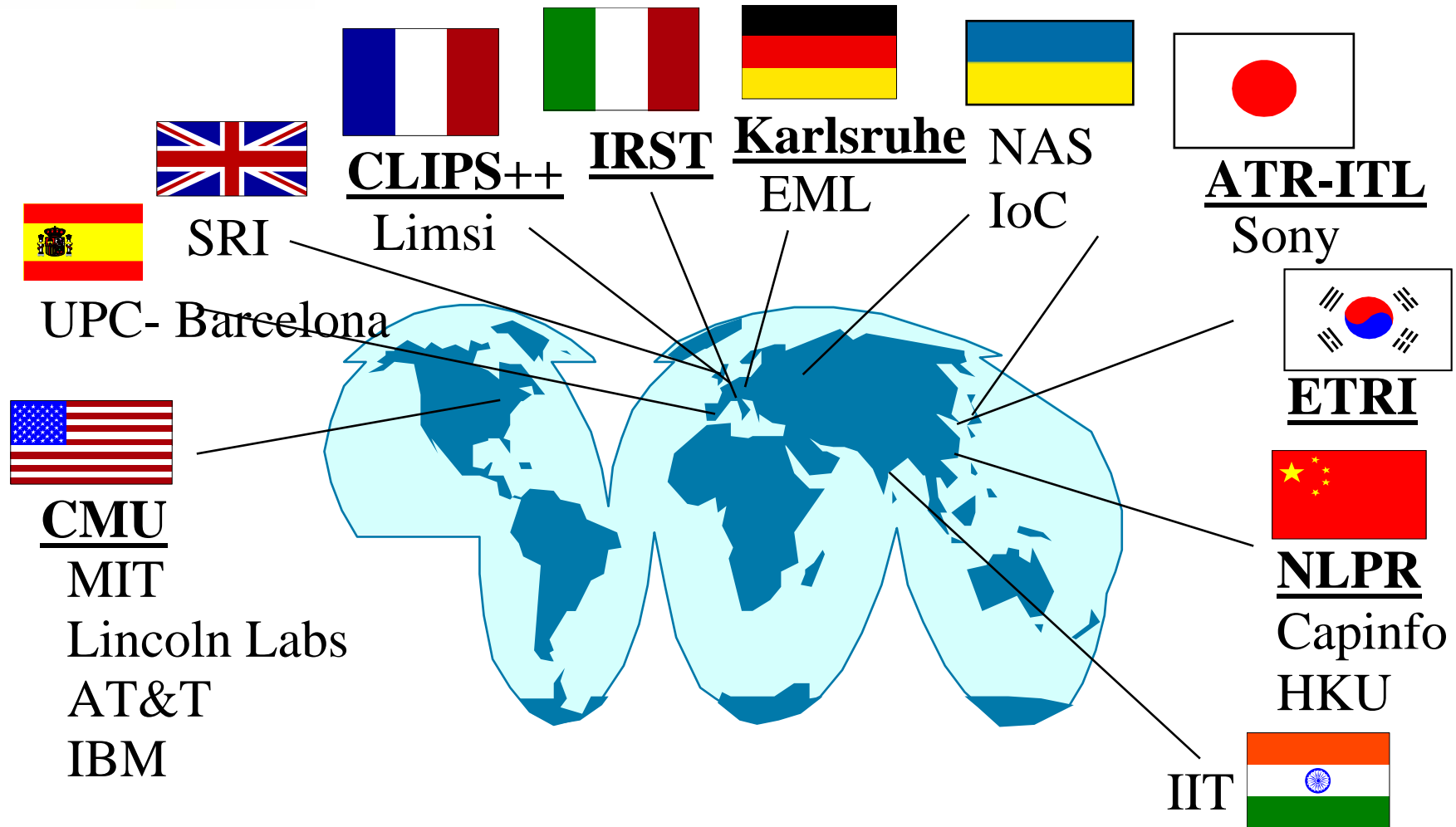
Speech to Speech Translation

Gianni Lazzari

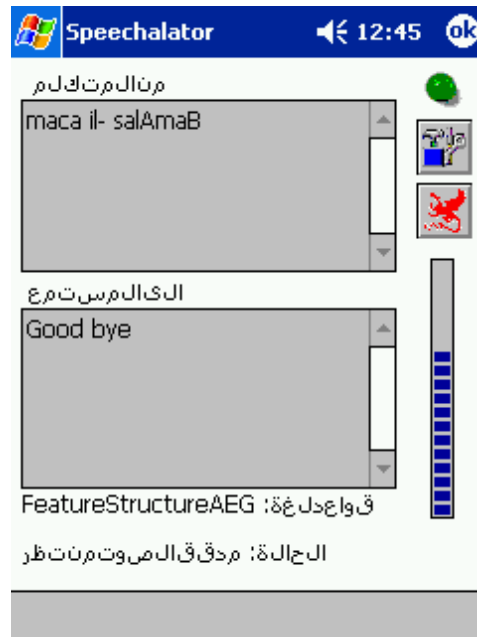
***HUMAN LANGUAGE TECHNOLOGY
EVALUATION***

Victoria Hotel Sliema Malta December 2

C-STAR, now



FACE TO FACE COMMUNICATION

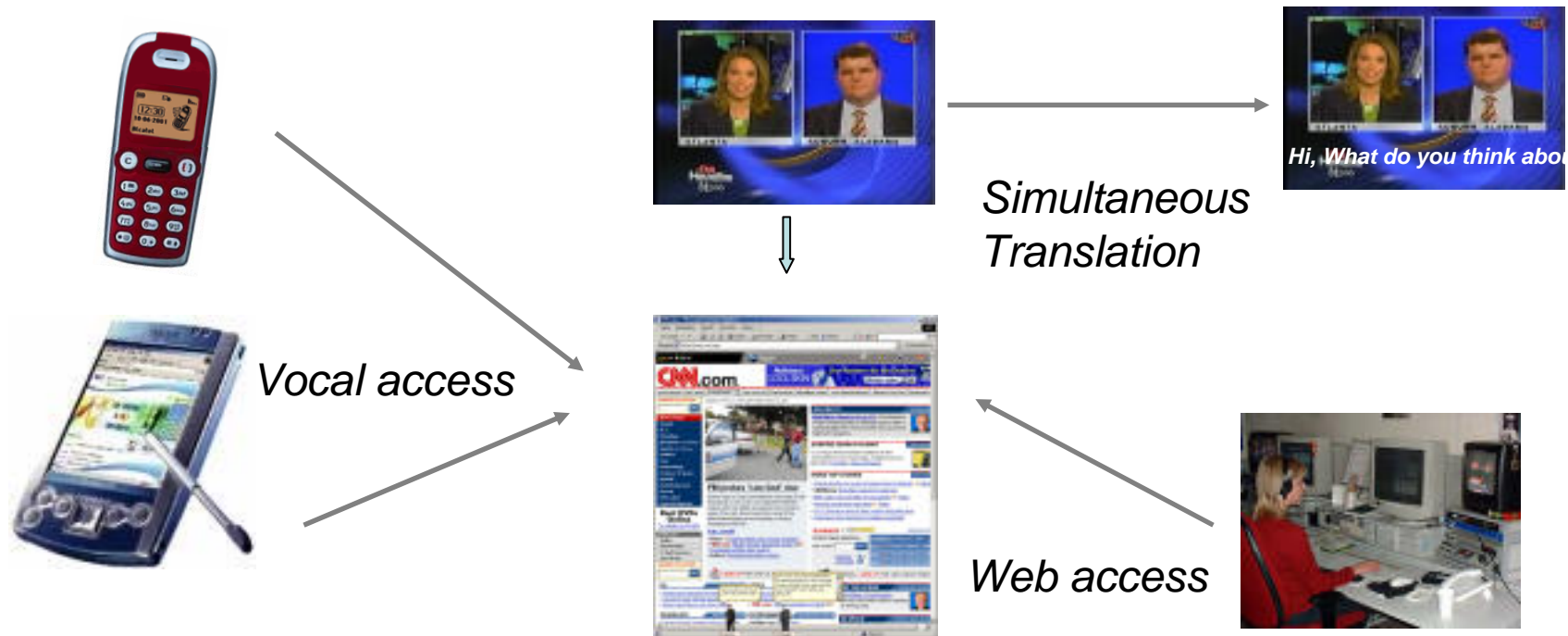


Translation of Lectures and Meetings



你们的评估准则是什么

Transcription and Translation of broadcast news, speeches and interviews



SST projects in the last 20 years



- **Pioneers**
 - C-STAR
 - IBM (statistical machine translation)
- **Demonstration oriented**
 - C-STAR II – VERBMOBIL - NESPOLE! -
BABYLON – DIGITAL OLIMPICS
- **Technology oriented**
 - C-STAR III (IWSLT)
 - TC-STAR
 - GALE (new US DARPA program)

Main Motivations for SST



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

Mainly promoted by Asian Countries
- **To let people understand news and content produced in foreign language**
 - Internet, Conferences, Multimedia documents

Mainly promoted by US for business and military

Mainly promoted by Europe. Member states have to preserve and promote their language, and through their language, their culture.

TC-STAR

Technology and Corpora for Speech to Speech Translation

moving...
from restricted domain
to unrestricted conversational speech
SLT

Contract Nr. FP6 506738

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



- 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
- Huge labor costs for transcription and translation
 - With 11 languages it was 549M Euros for translation.
- Recordings from *Europe by Satellite (EbS)*
 - Source language (speakers)
 - Target languages (interpreters)
- Texts from EU translation service



European Parliament audio data training

October 2005 status

detail			acoustic amount [h]		
			English EPPS	Spanish EPPS	Spanish PARL
total amount of recordings			177.3	172.6	44.2
transcribed speeches			91.5	61.8	38.3
male	interpreter	native	40.4	22.0	–
		non-native	0.9	1.8	–
	politician	native	11.0	8.8	27.2
		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	1.2	9.7
		non-native	0.6	0.3	–

Workplan



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

First Evaluation Campaign

April 2005

ASR & SLT

	Automatic Speech Recognition			Spoken Language Translation		
	EPPS EN	EPPS ES	VOA ZH	EN->ES	ES->EN	ZH->EN
IBM	X	X		X	X	X
IRST	X	X			X	X
LIMSI	X	X	X ²			
NOKIA	X					
RWTH	X	X		X	X	X
SONY						
UKA	X		X ²	X	X	X
UPC				X	X	
JHU ¹						X
UPV ¹				X	X	

Table 1: Participants in the First TC-STAR Evaluation Campaign

Moreover, we have computed the ES->EN score of the SYSTRAN system, to compare with those different systems.

Overview of the Campaign



- Evaluated Technologies: 2 out of 3
 - ASR
 - SLT
- Schedule: from 1 March 2005 to 15 April 2005 (longer than planned)
- Participants
 - 7 for ASR: 7 En, 4 Es, 2 Zh; no external total of 21 submissions
 - 5 for SLT + 2 external: 5 EnEs, 6 EsEn, 5 ZhEn → total 97 submissions

2 categories of tasks

- **EPPS**: English (En) and Spanish (Es), European Parliament Plenary Sessions
- **VoA**: Mandarin Chinese (Zh), Broadcast News from Voice of America (partly supplied by LDC)

ASR Tasks



- 2 Tasks
 - **EPPS:**
 - English 3 hours (~ 34 K words)
 - Spanish 3 h (~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 the cutoff date)

SLT Tasks



- 3 Tasks
 - **ASR**: translate automatic transcripts from ASR engines (ROVERed). No case, no punct.
 - *ASR-limsi*: translate output of Limsi's ASR, with case.
 - **Verbatim**: translate manual transcripts, with case, no punct.
 - **Text**: translate Final Text Edition (FTE) documents, with case and punct.
- 2 Conditions
 - **Primary**: use single-best hypo from ASR output, use only for training
 - EPPS: EPPS training set
 - VOA: LDC Large Data
 - **Secondary**: like primary plus ASR wordgraphs or any other optional input and publicly available data, and use any publicly available data for training

Schedule



- ASR
 - Development phase 4 Feb – 1 Mar 2005
 - ASR Run 2 Mar – 14 Mar 2005
 - Scoring phase:
 - EPPS : 20 Mar – 6 Apr
 - VoA : 10 Apr – 19 Apr
- SLT
 - Development phase: 15 Feb – 18 Mar 2005
 - SLT run: 21 Mar – 25 Mar 2005
 - Scoring phase:
 - EPPS: 28 Mar – 7 Apr 2005
 - VOA: 11 Apr – 15 Apr 2005

ASR Submissions En



	<i>Open</i>	<i>Public</i>	<i>Restricted</i>
IBM		1	1
IRST			4
LIMSI		1	1
NOKIA	5		
RWTH			5
SONY	1		
UKA		2	
Total	6	4	11



- *Es : 8 submissions in restricted conditions*
 - *IBM (1)*
 - *IRST (2)*
 - *LIMSI (1)*
 - *RWTH (4)*
- *Zh : 1 common submission from LIMSI/UKA*

SLT Participants - EnEs



Direction	Set	Condition	Organisation	Comments
En->Es	ASR	Primary	IBM	
En->Es	ASR	Secondary	IBM	2 submissions
En->Es	ASR	Secondary	RWTH	single-best + wordgraphs
En->Es	ASR	Primary	RWTH	2 submissions
En->Es	ASR	Primary	UKA	
En->Es	ASR	Primary	UPV	
En->Es	ASR-limsi	Primary	UPC	
En->Es	text	Primary	IBM	2 submissions
En->Es	text	Secondary	IBM	2 submissions
En->Es	text	Primary	RWTH	4 submissions
En->Es	text	Primary	UKA	
En->Es	text	Primary	UPC	
En->Es	text	Primary	UPV	
En->Es	verbatim	Primary	IBM	
En->Es	verbatim	Secondary	IBM	2 submissions
En->Es	verbatim	Primary	RWTH	2 submissions
En->Es	verbatim	Primary	UKA	
En->Es	verbatim	Primary	UPC	

Total: 28 submissions



Information Society

SLT Participants - EsEn



Direction	Set	Condition	Organisation	Comments
Es->En	ASR	Primary	IBM	2 submissions
Es->En	ASR	Secondary	IBM	2 submissions
Es->En	ASR	Primary	ITC	3 submissions
Es->En	ASR	Secondary	RWTH	single-best + wordgraphs
Es->En	ASR	Primary	RWTH	2 submissions
Es->En	ASR	Primary	UKA	
Es->En	ASR	Secondary	UKA	
Es->En	ASR	Primary	UPV	
Es->En	ASR-limsi	Primary	UPC	
Es->En	text	Primary	IBM	3 submissions
Es->En	text	Secondary	IBM	3 submissions
Es->En	text	Primary	ITC	2 submissions
Es->En	text	Primary	RWTH	2 submissions
Es->En	text	Primary	UKA	
Es->En	text	Primary	UPC	
Es->En	text	Primary	UPV	
Es->En	verbatim	Primary	IBM	2 submissions
Es->En	verbatim	Secondary	IBM	2 submissions
Es->En	verbatim	Primary	ITC	2 submissions
Es->En	verbatim	Primary	RWTH	2 submissions
Es->En	verbatim	Primary	UKA	
Es->En	verbatim	Primary	UPC	

Total: 38 submissions

SLT Participants - ZhEn



Direction	Set	Condition	Organisation	Comments
Zh->En	ASR	Primary	IBM	4 submissions
Zh->En	ASR	Primary	ITC	2 submissions
Zh->En	ASR	Primary	JHU	
Zh->En	ASR	Secondary	JHU	
Zh->En	ASR	Primary	RWTH	
Zh->En	ASR	Primary	UKA	
Zh->En	text	Primary	IBM	5 submissions
Zh->En	text	Primary	ITC	2 submissions
Zh->En	text	Primary	JHU	
Zh->En	text	Secondary	JHU	
Zh->En	text	Primary	RWTH	
Zh->En	text	Primary	UKA	
Zh->En	verbatim	Primary	IBM	4 submissions
Zh->En	verbatim	Primary	ITC	2 submissions
Zh->En	verbatim	Primary	JHU	
Zh->En	verbatim	Secondary	JHU	
Zh->En	verbatim	Primary	RWTH	
Zh->En	verbatim	Primary	UKA	

Total: 31 submissions

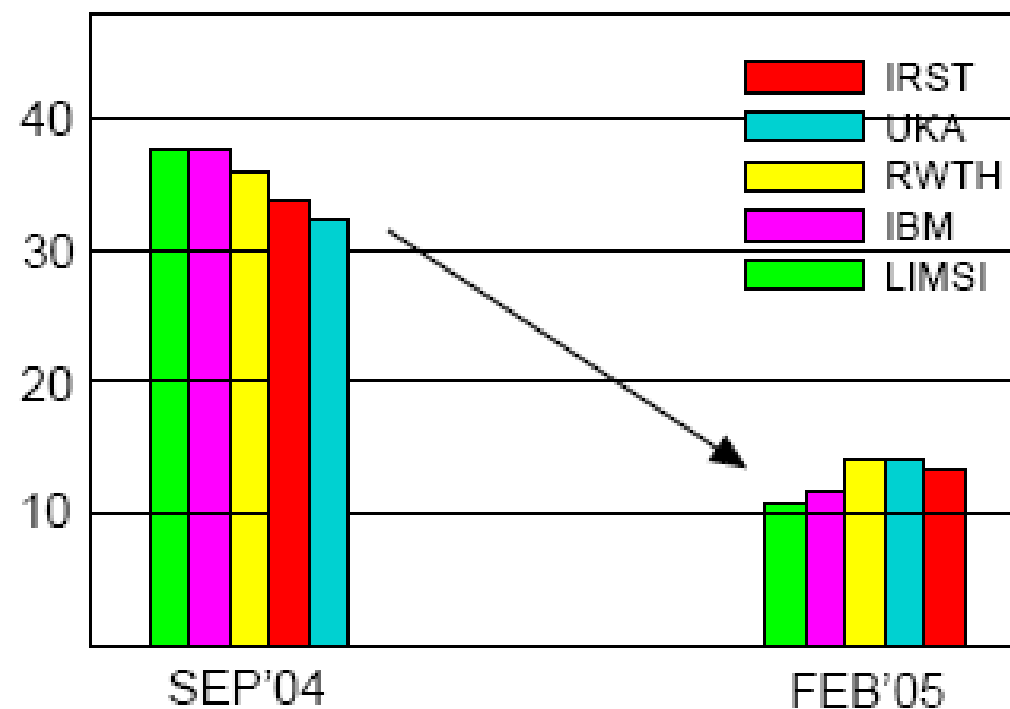
Resources



- Development and test resources taken from the same original sets for ASR and SLT
- EPPS (English, Spanish): data from European Parliament satellite broadcast, usage rights negotiated
 - ASR: audio files + manual transcripts ~4h
 - SLT: subset of manual transcripts ~25000 words + 25000 words taken from corresponding FTE documents
- VOA (Mandarin): original data available at LDC, rights must be acquired by participants
 - ASR: audio files + resegmented and corrected manual transcripts ~3h
 - SLT: subset of resegmented manual transcripts ~15000 words

ASR results

Progress on English EPPS



EPPS English: 32.0 → 10.6 (-67%) [Rover: 9.5]

ASR results



- Quite succesful at reducing the WER, from 30% to 10% : ASR on EPPS (English): **10% WER**
- MT systems can use lower WER output (especially as MT models get better)
- Still plenty of room for improvment at all the levels
- More data, more collaboration and better combination (system cascade and rover with Zh)



Some ASR open issues



- **Manual vs. automatic segmentation**
 - ASR does not need manual segmentation
 - Segmentation is needed to produce lattice for MT
- **Rich transcription**
 - Confidence scores (should be mandatory)
 - Case sensitive output
 - Punctuation, sentence breaks
- **Lattices and confusion networks**

- **Evaluation Measures**

automatic measures based on single or multiple reference translations:

- **WER = Word Error Rate:**

- Levenshtein (edit) distance (as in speech recognition)

- **PER = Position independent word Error Rate:**

- ignore word order and count word errors

- **BLEU = 'Bilingual Evaluation Understudy'**

- accuracy measure: geometric mean of n-gram precision + brevity penalty

- **NIST = NIST variant of BLEU**

- accuracy measure: arithmetic mean of n-gram precision + brevity penalty

remark: these automatic measures correlate with human judgement (adequacy + fluency)

Evaluation Results EN→SP



Input	Site	BLEU [%]	NIST	PER [%]	WER [%]
ASR (WER= 9.5%)	RWTH	38.7	8.73	38.6	49.8
	IBM	34.3	8.13	42.0	54.5
	UPC*	33.8	8.00	43.1	54.0
	UKA	33.0	7.94	43.4	55.9
	UPV	19.1	5.46	53.3	62.5
Verbatim	RWTH	42.5	9.32	35.4	46.1
	UPC	38.1	8.72	39.2	49.5
	IBM	36.8	8.55	39.6	51.8
	UKA	33.4	8.29	41.8	53.2
Text	UPC	46.2	9.65	32.7	41.2
	IBM	45.2	9.44	32.8	43.2
	RWTH	38.9	8.72	36.1	48.4
	UKA	37.6	8.46	38.3	49.6
	UPV	34.1	7.51	40.8	48.7

external participant: UPV

- strong correlation between the four measures
- difference between WER and PER: 10-15%
- only small degradation:
 - from text to verbatim input
 - from verbatim to ASR input

Evaluation Results SP→EN



Input	Site	BLEU [%]	NIST	PER [%]	WER [%]
ASR (WER= 10.1%)	RWTH	41.5	9.12	35.4	46.6
	IBM	39.7	8.81	37.7	48.6
	UPC*	37.7	8.56	39.2	48.7
	ITC	34.7	7.97	42.8	53.8
	UKA	32.3	7.85	43.1	55.0
	UPV	16.0	4.35	57.1	63.6
Verbatim	RWTH	45.9	9.75	31.7	42.5
	IBM	44.1	9.47	33.4	43.9
	UPC	42.1	9.26	34.9	44.1
	ITC	38.1	8.46	39.8	50.0
	UKA	33.4	7.96	43.3	54.5
Text	UPC	53.3	10.55	27.1	35.1
	IBM	53.1	10.38	27.0	35.9
	ITC	47.5	9.60	31.3	40.6
	RWTH	46.1	9.68	29.7	40.5
	UKA	40.5	8.96	34.4	44.8
	UPV	32.7	6.80	41.3	47.5

- compare with English-to-Spanish:
- similar observations
 - better absolute performance

Evaluation Results Zh→EN

Input	Site	BLEU [%]	NIST	PER [%]	WER [%]
ASR (CER \approx 9.5 %)	RWTH	16.2	5.87	57.8	78.1
	UKA	13.5	5.46	61.8	81.2
	JHU	13.2	5.43	63.8	85.3
	IRST	11.5	5.20	63.6	83.7
	IBM	5.2	2.61	90.0	104.5
Verbatim	RWTH	16.8	5.99	58.0	78.6
	IBM	13.7	5.70	62.4	86.6
	UKA	13.6	5.64	60.8	80.8
	JHU	13.4	5.58	63.1	84.9
	IRST	12.0	5.37	62.8	83.6
Text	RWTH	16.5	5.95	55.4	75.8
	JHU	14.6	5.75	58.9	80.8
	UKA	14.2	5.67	58.2	78.2
	IBM	13.9	5.67	60.9	84.8
	IRST	12.6	5.36	61.4	82.3

external participant: JHU

- correlation between the four measures
- difference between WER and PER: 20-25%
- only small degradation:
 - from text to verbatim input
 - from verbatim to ASR input
- significantly worse performance than for EPPS tasks ??

Example Spanish to english



VERBATIM	los proyectos de enmienda deberán presentarse con la firma de treinta y siete diputados como mínimo o en nombre de una Comisión
ASR	los proyectos de enmienda deberán presentarse con la firma del treinta y siete diputados como mínimo buen nombre de una comisión
TRANS VERBATIM	the amendment of projects must be made with the signing of thirty seven Members as minimum or on behalf of a committee
TRANS ASR	the amendment of projects must be made with the signing of the thirty seven Members as minimum good name of a Commission
REFERENCE 1	the draft amendments must be tabled with the signatures of at least thirty-seven Members or on behalf of a Committee
REFERENCE 2	amendment projects must be presented with the signature of at least thirty-seven members or on behalf of a Committee

SLT results



- **Operational systems:**
 - three language pairs (C-E, S-E, E-S)
 - different types of input: ASR, verbatim, text
 - a couple of different SLT systems
- **Unconstrained Speech:**
 - real-life tasks: EPPS for Spanish-to-English and English-to-Spanish (availability of EPPS data!)
 - evaluation: full ASR-SLT system with 2 external participants
- **Surprisingly good performance for EPPS tasks**
 - variability of 20% relative across different SLT systems
 - Joint ASR + *MT* : **35% PER**
 - worldwide first systems for real-life speech translation and first evaluation campaign

Some research topics under investigation



- **Spoken Language Translation**
 - How to introduce linguistic knowledge in the statistical MT approach ?
 - How to improve and/or introduce new innovative methods ?
- **Interfaces**
 - ASR-> SLT
 - SLT -> TTS
- **SLT and SST evaluation**
 - automatic scoring
 - human evaluation

Innovative Methods.

- Improvement of the technology in alignment, lexicon and language models.
- Extension of the models to include syntactic knowledge and structures
- ROVER-like and other combinations of SLT outputs

Human-Supplied Knowledge.

- Use of bilingual dictionaries, POS tagging, and morpho-syntactic analysis in the statistical approach (e.g. preprocessing, postprocessing, integrated approach)

- **ASR → SLT**
 - from single best to word graph, N-best
 - Definition of a wordgraph interface between ASR and SLT and its use in the translation process
 - How to manage speech disfluencies
- **SLT → TTS**

Challenge:

 - Pass on characteristics of the speech signal from source to target language in addition to text generated by SLT
 - Approach: separate synthesis from translation problem

Various types of information:

- ***text as the SLT output.***

conventional interface in text-to-speech synthesis

- ***additional output of SLT:***

- confidence measures for each target word (phrase)

- information in the spoken source language:

- characteristics of the speaker ('voice conversion')

- intonation and prosody

- phrase boundary markers

- disfluencies, hesitations, ...

- ...

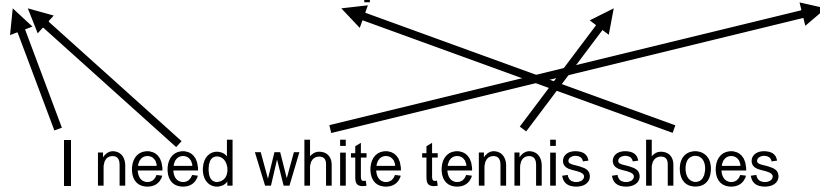
Issue:

- acoustic analysis of the spoken source language is needed

- unclear: exactly what type of information is needed?

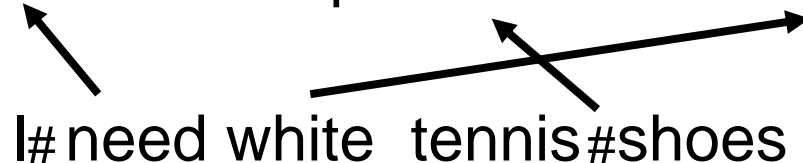
From word-based SMT...

Vorrei delle scarpe da tennis bianche



... to phrase-based S

Vorrei#delle scarpe#da#tennis bianche





single- word /
phrase-based
statistical lexicon:

- word inflection

- model training:

- translation:

Und Sie rufen mich an ~ and you will call me

... und rufe Sie an

... and ??? you me

Sie ~ you

rufen ~ call
geben ~ give

mich ~ me
mir ~ me

rufen#mich#an ~ call#me
Ihre ~ your

rufen ~ will#call

Sie#geben ~ you#give

geben#mir ~ give#me

mir#Ihre ~ me#your

Telefonnummer ~ telephone#number

- differences in word order

- training:

Sie geben mir Ihre Telefonnummer ~
You give me your telephone number

- test:

... falls Sie mir Ihre Telefonnummer geben

... if you me#your telephone#number give

- use morphological information in translation lexicon
- use syntax information for source sentence re-ordering

... und rufe Sie an

ruf|e ~ call
ruf|st ~ call
ruf|en ~ call

... falls Sie mir Ihre Telefonnummer geben

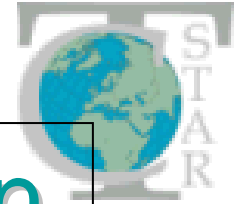
... falls <pers.pronoun> mir Ihre Telefonnummer <verb>

... falls Sie geben mir Ihre Telefonnummer

... if you#give#me#your#telephone#number



Information Society
Technologies



Improving Speech Translation

- Translation of 1st Best recognizer output
- Improvements:
 - Using N-Best Recognizer output
 - Using Speech Recognizer Lattice
 - Preprocessing to remove disfluencies

1-Best Input

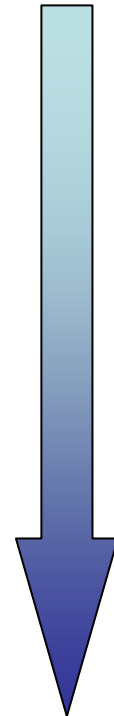
Speech Recognition

First Best Recognizer Output



Machine Translation

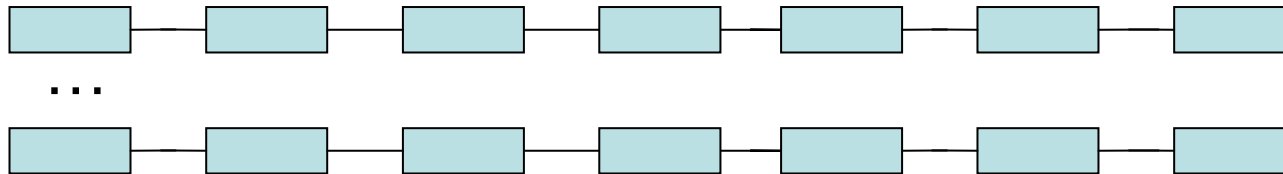
First Best Translation



N-Best Input

Speech Recognition

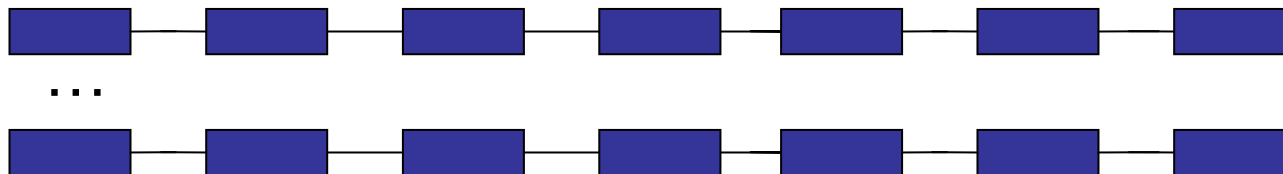
N-Best Recognizer Outputs



Machine Translation

N-Best Translation

Possible Reranking!

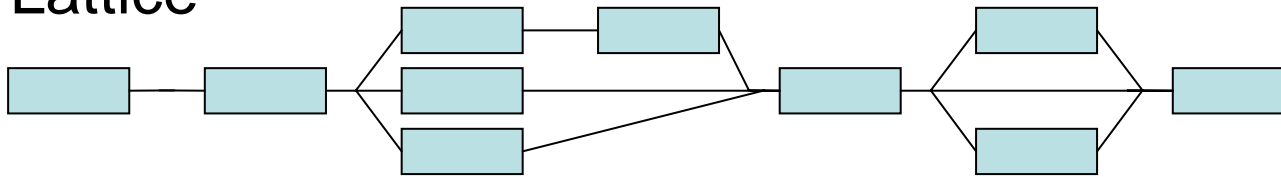


Tight Coupling

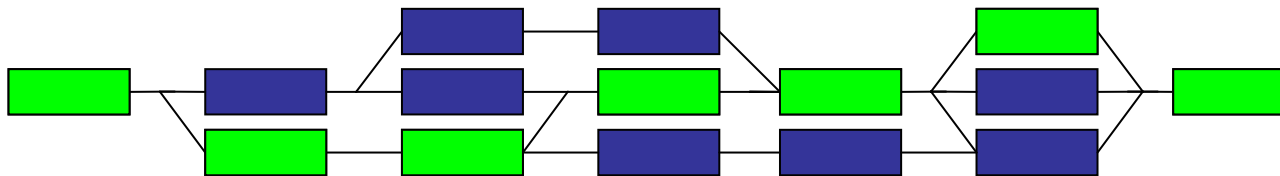


Speech Recognition

Lattice



Machine Translation



Translated Lattice

Best Path

- other issues:
 - effect of ASR/SLT errors on synthesis
 - synchronization between speech in source and target language: segmentation, word order, ...

SLT evaluation



- .. *Nothing .. is perfect...*
 - **Bleu Nist Per Wer Gtm Meteor Ter**
 - **Adequacy Fluency Meaning Maintenance**.....and now “**Edit Distance**”(Gale)

*we want a consistent & fast progress
any metric that help to measure any new
idea is welcome... if it is also simple to
communicate to users then is better...*



SLT evaluation



5 TC-star partners participate in the last IWSLT 2005 evaluation campaign and workshop

Overview of the IWSLT 2005 Evaluation Campaign

Matthias Eck and Chiori Hori

Interactive Systems Laboratories Carnegie Mellon University {matteck, chiori}@cs.cmu.edu

Subjective Evaluation – Fluency/Adequacy



Fluency		Adequacy	
4	Flawless English	4	All information
3	Good English	3	Most information
2	Non-Native English	2	Much information
1	Disfluent English	1	Little information
0	Incomprehensible	0	None

Typically used metrics Fluency/Adequacy
(e.g. IWSLT 2004)

Here: **0 – 4** instead of **1 – 5**

Subjective Evaluation Meaning Maintenance

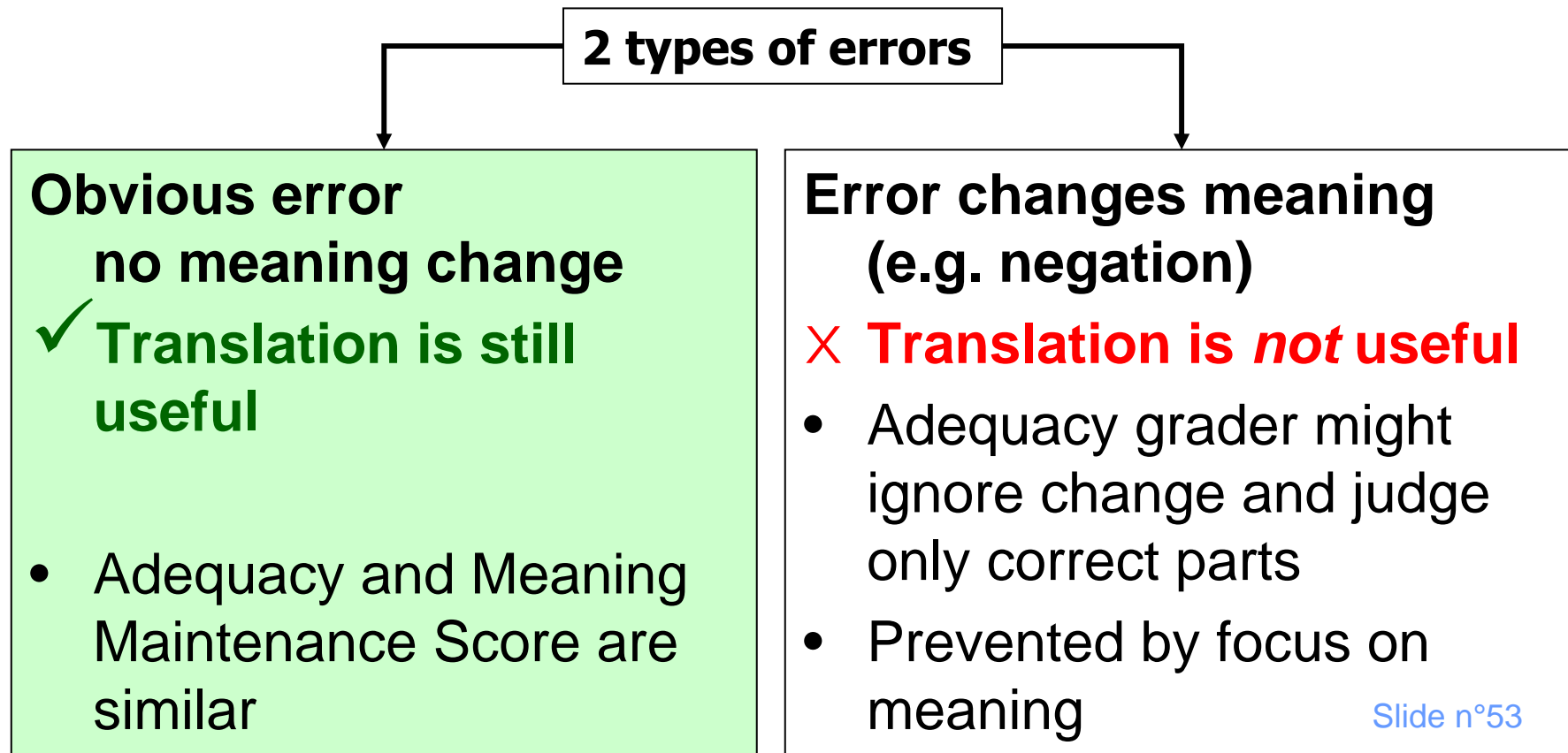
Meaning Maintenance	
4	Exactly the same meaning
3	Almost the same meaning
2	Partially the same meaning and no new information
1	Partially the same meaning but misleading information is introduced
0	Totally different meaning

Adequacy	
4	All information
3	<i>Most</i> information
2	<i>Much</i> information
1	<i>Little</i> information
0	None

Why Meaning Maintenance?



- Focus on comparing meaning of translation with source
- Degree of misleading information?



- All translations shown at the same time
 - Randomly ordered
 - Comparison among translations of the same sentence
- No explicit reference
 - reference included in translations
 - No bias by shown reference
 - gives oracle score
- Source is shown for Adequacy and Meaning Maintenance scores
- 5 bilingual graders (scores shown are for 3 graders)
- First all Fluency scores, then Adequacy, finally Meaning Maintenance



Human Evaluation Results



Adequacy	
MIT-LL/AFRL	2.71
ITC-IRST	2.65
RWTH	2.63
TALP-phrase	2.52
IBM	2.51
TALP-ngram	2.44
EDINBURGH	2.33
ATR-C3	2.31
NTT	2.09
CMU	1.95
USC-ISI	1.90

Fluency	
ITC-IRST	3.15
RWTH	3.04
CMU	2.88
ATR-C3	2.86
TALP-ngram	2.82
EDINBURGH	2.81
MIT-LL/AFRL	2.79
TALP-phrase	2.78
IBM	2.77
USC-ISI	2.32
NTT	1.97

Mean. Maint.	
MIT-LL/AFRL	2.63
RWTH	2.60
ITC-IRST	2.60
TALP-phrase	2.49
IBM	2.44
TALP-ngram	2.40
EDINBURGH	2.35
ATR-C3	2.23
NTT	2.03
USC-ISI	1.96
CMU	1.94

Consistency? (Inter - Grader)



- How consistent are the scores assigned by the 3 graders?
- Average differences between grades:
 - Agreement between all 3 graders for about 40% of sentences
 - Agreement between 2 graders for about 60% of sentences

	Adequacy	Fluency	Mean. Maint.
G1-G2	0.58	0.54	0.54
G1-G3	0.70	0.69	0.93
G2-G3	0.60	0.75	0.80
AVG	0.63	0.66	0.76

Consistency? (Intra – Grader)



- How consistent are the scores assigned by each grader?
(based on 10% sentences graded twice)
- Average differences between first and second grade:

	Adequacy	Fluency	Mean. Maint.
Grader 1	0.32	0.29	0.25
Grader 2	0.32	0.30	0.24
Grader 3	0.60	0.61	0.40
AVG	0.41	0.40	0.30

Translation target



- Manual transcription (plain sentences in BTEC)
- ASR output of spoken BTEC sentences

Where would you like to go?

Is there a discount for children?

Did you have fun today?

Sure. Can I have a receipt?

I'd like to try some local wine.

- No discourse.

Chinese ⇔ English – Supplied Data - Rankings

BLEU	NIST	mWER	mPER	GTM	METEOR	TER	Adeq.	Fluency	Mean. M.
ITC-IRST	RWTH	ITC-IRST	ITC-IRST	ITC-IRST	MIT/AFR	ITC-IRST	MIT/AFR	ITC-IRST	MIT/AFR
RWTH	MIT/AFR	RWTH	MIT/AFR	MIT/AFR	ITC-IRST	EDINBG	ITC-IRST	RWTH	RWTH
EDINBG	ITC-IRST	EDINBG	RWTH	TALP-ph	RWTH	RWTH	RWTH	CMU	ITC-IRST
TALP-ph	IBM	TALP-ph	TALP-ph	RWTH	TALP-ph	TALP-ph	TALP-ph	ATR-C3	TALP-ph
MIT/AFR	TALP-ng	MIT/AFR	IBM	EDINBG	TALP-ng	MIT/AFR	IBM	TALP-ng	IBM
TALP-ng	ATR-C3	IBM	EDINBG	IBM	IBM	IBM	TALP-ng	EDINBG	TALP-ng
CMU	TALP-ph	TALP-ng	TALP-ng	TALP-ng	EDINBG	TALP-ng	EDINBG	MIT/AFR	EDINBG
IBM	NTT	CMU	ATR-C3	ATR-C3	ATR-C3	CMU	ATR-C3	TALP-ph	ATR-C3
ATR-C3	EDINBG	ATR-C3	CMU	USC-ISI	NTT	ATR-C3	NTT	IBM	NTT
USC-ISI	CMU	USC-ISI	USC-ISI	CMU	USC-ISI	USC-ISI	CMU	USC-ISI	USC-ISI
NTT	USC-ISI	NTT	NTT	NTT	CMU	NTT	USC-ISI	NTT	CMU

A new metric TER



TER – Translation Error Rate

0 1

Newly introduced metric:

- Measure error as the minimum number of edits needed to change hypothesis so that it exactly matches one of the references
 - $TER = \langle \# \text{ of edits} \rangle / \langle \text{avg} \# \text{ of reference words} \rangle$
 - TER is calculated against best (closest) reference
- Edits include insertions, deletions, substitutions and shifts
 - All edits count as 1 error (=edit distance)
 - Shift moves a sequence of words within the hypothesis
 - Shift of any sequence of words (any distance) is only 1 error

Plan



1. Overview of the Campaign
2. Evaluation Tasks, Schedule & Participants
3. Production of Resources
4. Validation of Resources
5. ASR Evaluation
6. SLT Evaluation
7. Conclusions



<u>Training</u>	
EPPS	40h audio + manual transcripts from sessions recorded 3 May – 14 Oct 2004 → produced by RWTH (En) and UPC (Es)
VOA	publicly available sources + VOA 1998 available at LDC without Dec. 1998 (audio + LDC transcripts)
Development	
EPPS	4h audio + manual transcripts (sessions 25-28 Oct 2004) → produced by ELDA (En+Es)
VOA	3h audio + resegmented manual transcripts (1-11 Dec 1998) → transcripts by LDC resegmented by ELDA
Test	
EPPS	4h audio + manual transcripts (sessions 15-18 Nov 2004) → produced by ELDA (En+Es)
VOA	3h audio + resegmented manual transcripts (14-22 Dec 1998) → transcripts by LDC resegmented by ELDA



<u>Training</u>	
EPPS	En+Es manual transcripts from ASR training (RWTH and UPC) + FTE April 1996 – 14 Oct 2004
VOA	publicly available sources (bilingual and monolingual)
Development	
EPPS	manual transcripts (ELDA) taken from ASR development ~25000 words + 2 reference translations (ELDA)
VOA	manual transcripts (LDC) resegmented (ELDA) of 1h30 excerpt from ASR development (1-3 Dec 1998) ~14000 words + 2 English reference translations (ELDA)
Test	
EPPS	manual transcripts (ELDA) taken from ASR test ~25000 words + 2 reference translations (ELDA)
VOA	manual transcripts (LDC) resegmented (ELDA) of 1h30 excerpt from ASR test (14-16 Dec 1998) ~15000 words + 2 English reference translations (ELDA)



- Development Set
 - Transcripts En and Es come from the same portions of audio files → they are translations of each other
 - En and Es: very low proportion (5%) of En/Es speaking politicians, many interpreters
 - Transcripts were artificially resegmented to match SLT requirements
 - Quality of reference translations is not as good as expected
- Test Set
 - Subsets of the transcripts were chosen to collect all En/Es speaking politicians (>> 50%) + interpreters up to 25kw → they are different subsets
 - Transcripts are well segmented from the beginning
 - Quality of reference translations is better
 - One of ZhEn reference translations was delayed

Validation tasks



- Transcription validation for
 - Dev.set and test set of
 - English
 - Spanish
 - 2000 segments
- Translation validation for
 - Dev.set and test set of
 - English -> Spanish
 - Spanish -> English
 - Mandarin -> English
 - 1200 words per source text



Validation Remarks: Transcriptions



- Errors computed at segment level
- No events in test sets
- Test set English contained 1 Spanish file

Validation Results: Transcriptions

Transcription val.	Crite rion	Dev.		Test	
		EN	ES	EN	ES
%Speech	5	6.7	1.7	2.5	0.5
%Non- speech	10	2.3	0.4	NA	NA
%Speaker	2	0.0	0.0	0.1	0.0
%Segment	5	1.3	1.3	1.0	4.0
%Lextag	5	0.0	0.0	NA	NA

Validation Remarks: Translations



- 1200 words from contiguous segments from source text (except Mandarin; there from target text)
- Two translations per text from different agencies

- Scoring:

Error	Penalty
Syntactical	4
Deviation from guidelines	3
Lexical	2
Poor usage	1
Punctuation/spelling	0.5 (max 10)

- Max. 40 penalty points per translation allowed

Validation Results: Translations

	Dev.		Test	
	Trans1	Trans2	Trans1	Trans2
En-Es	41	81	45.5	65
Es-En	25.5	88	10.5	58.5
Zh-En	156.5	97.5	Under val.	Under val.

Validation Conclusions



- Not all data was of sufficient quality
- Validation was circumstantial since data were in full use. Validation should be given more time in a next evaluation round
- New evaluation round with updated data
- How about validation of train sets
 - Will they be distributed to third parties?



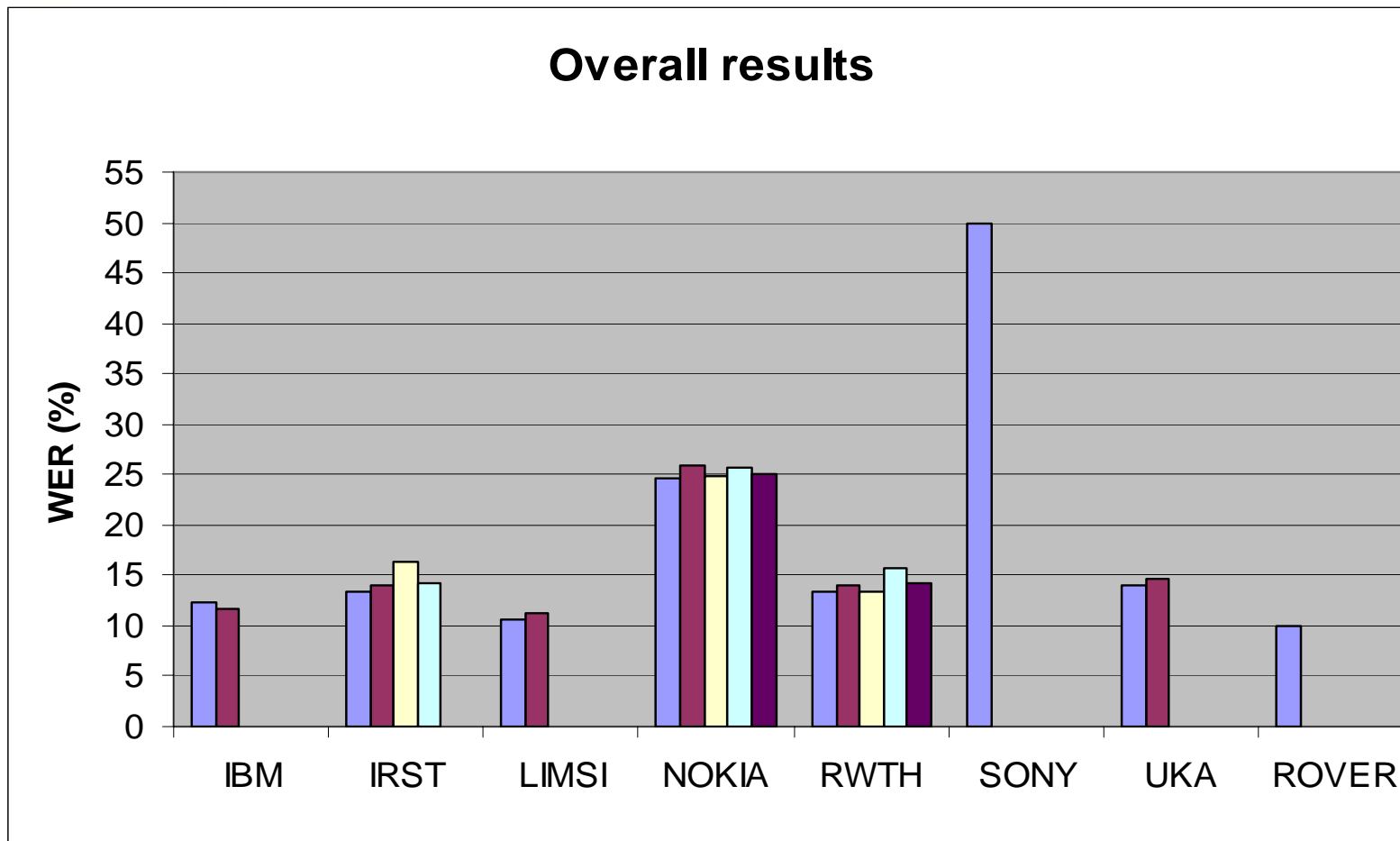
Evaluation Results

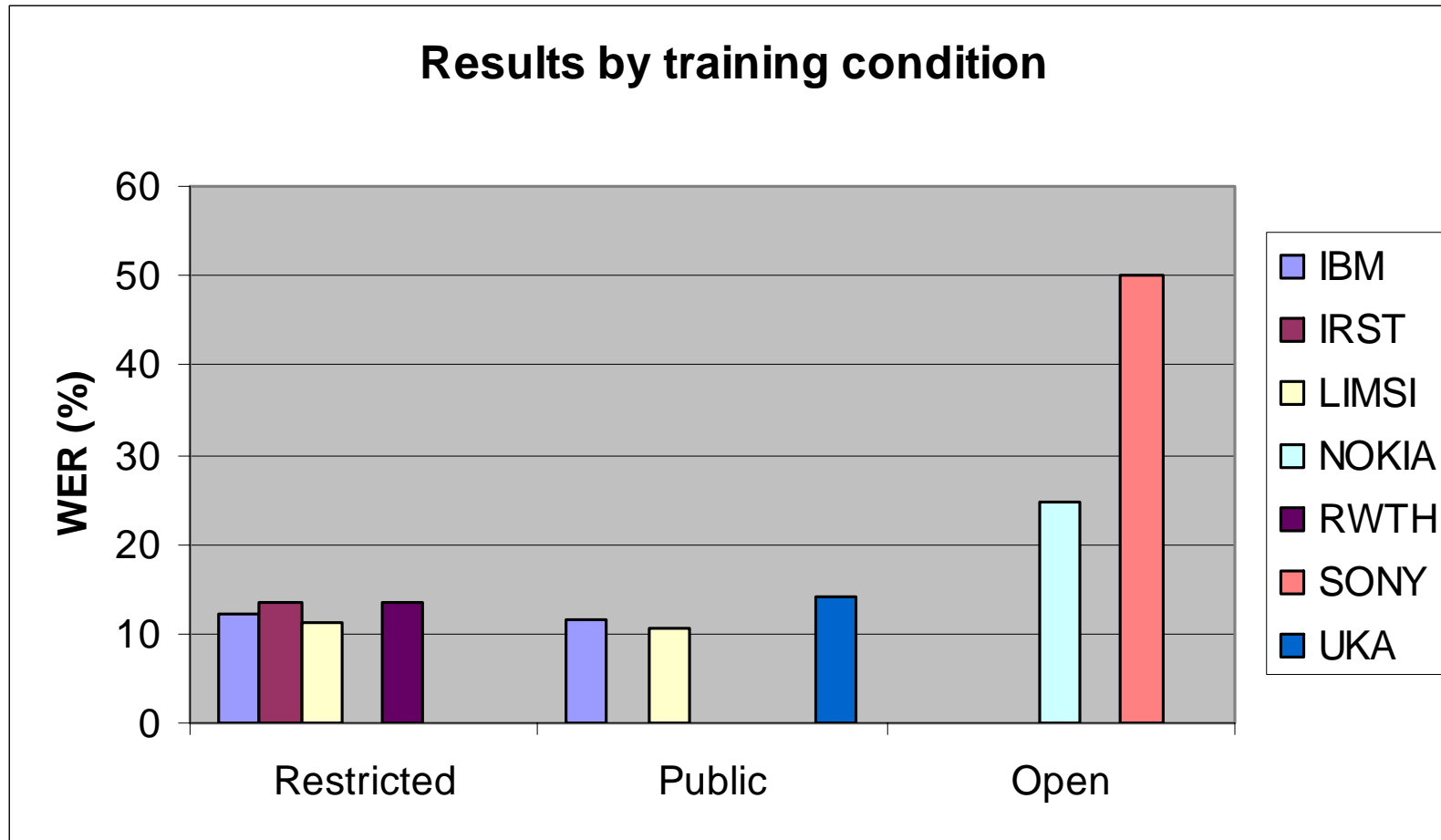
ASR Results En (1/3)



	<i>Open</i>	<i>Public</i>	<i>Restricted</i>
IBM		11.6	12.3
IRST			13.4
LIMSI		10.6	11.2
NOKIA	24.6		
RWTH			14.1
SONY	50.0		
UKA		14.0 (13.7*)	
ROVER			9.9

Results of primary systems in WER



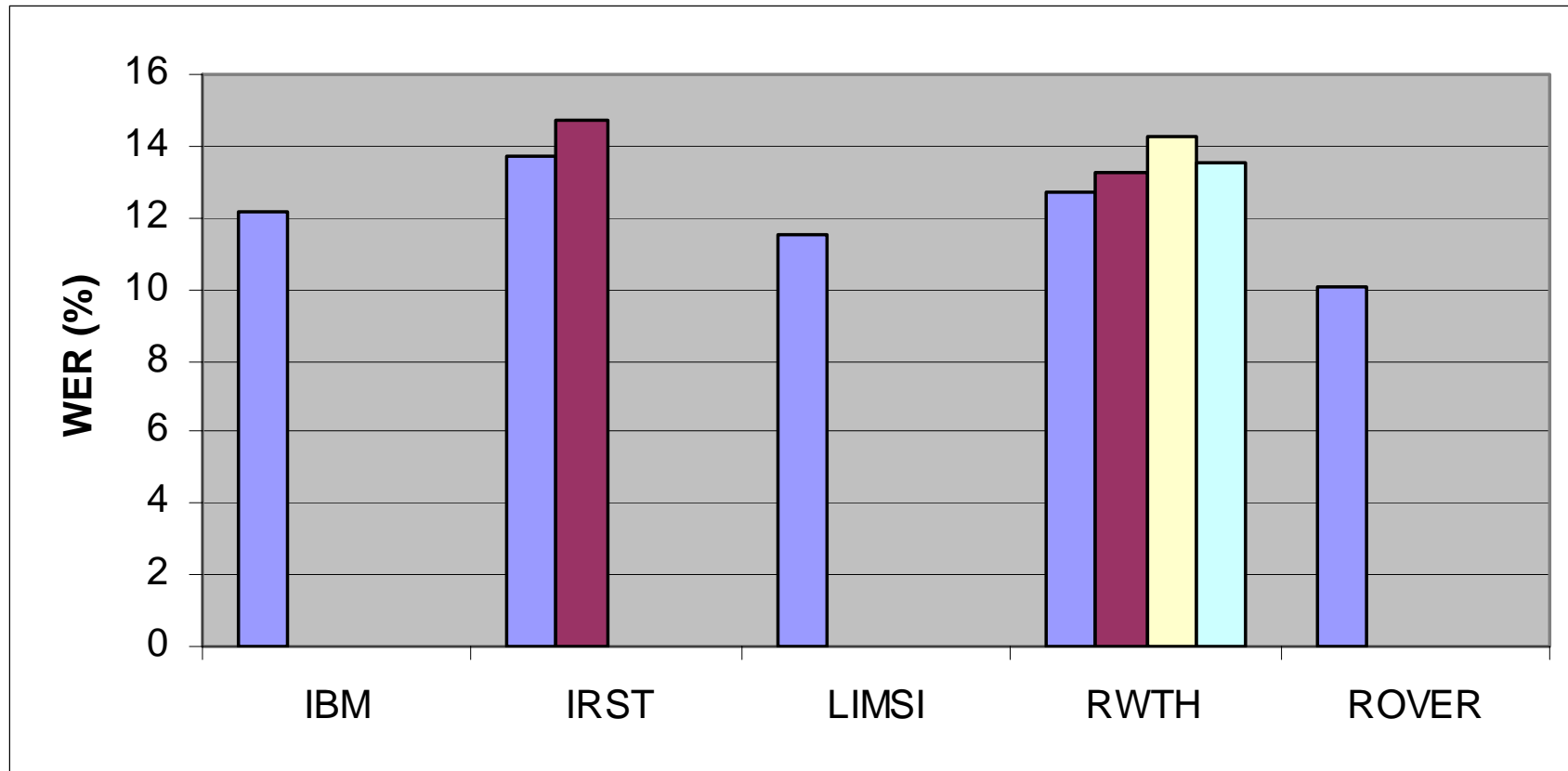


ASR Results Es (1/2)



	<i>Restricted</i>
IBM	12.2
IRST	13.7
LIMSI	11.5
RWTH	12.7
<i>ROVER</i>	<i>10.1</i>

Results of primary systems in WER





ASR Results Zh



- 1 common submission from LIMSI-UKA
- WER <10 %
- We are fixing some errors in the reference

SLT Evaluation



- For each task and condition, more than 1 submission allowed per participant
- Scoring
 - Metrics: BLEU/NIST v11, mWER, mPER, mCER (for ZhEn)
 - Scoring tools developed by ELDA (except BLEU/NIST)
 - Other metrics will be available later on (X-Score, D-Score)
- Human evaluation? To be set up.

BLEU



- **BLEU**

- Geometric mean of n-gram precision of hypothesis compared to the reference translation
- Length Penalty for short translations

- **Benefits**

- Missing references can be covered by combining of other references
- Correlates well with Fluency

- **Scores:**



- **Problems**

- **Re-combination** of references could cause errors
- All words are equally important
- Weak correlation with Adequacy

- **NIST**
 - Variant of BLEU using arithmetic mean of weighted n-gram precision values
- **Benefits**
 - Considers information gain
 - Up to 9-grams, usually 5-grams
 - Good correlation with Adequacy

- **Scores:**



- **Problems**
 - **Re-combination** of references could cause errors
 - Weak correlation with Fluency (human judgement)

mWER, mPER



- **mWER**
 - Word Error Rate on multiple references
 - edit distance:
hypothesis \Leftrightarrow closest reference
- **mPER**
 - mWER without considering word order
- **Benefits**
 - Correlates well with human judgement...

- **Scores:**



- **Problems**
 - ...if enough references are available

GTM, METEOR



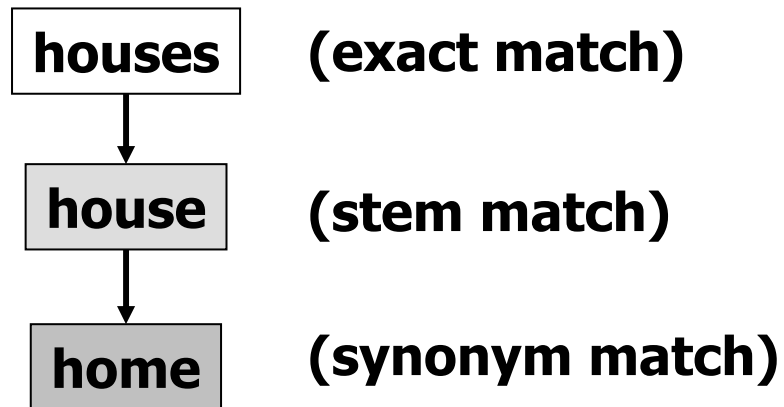
- **GTM**

- Similarity between texts using unigram based F-measure

- **METEOR**

- Considers:
 - ▶ Exact matches
 - ▶ stem matches
 - ▶ synonym matches (using WordNet)

- **Scores:**

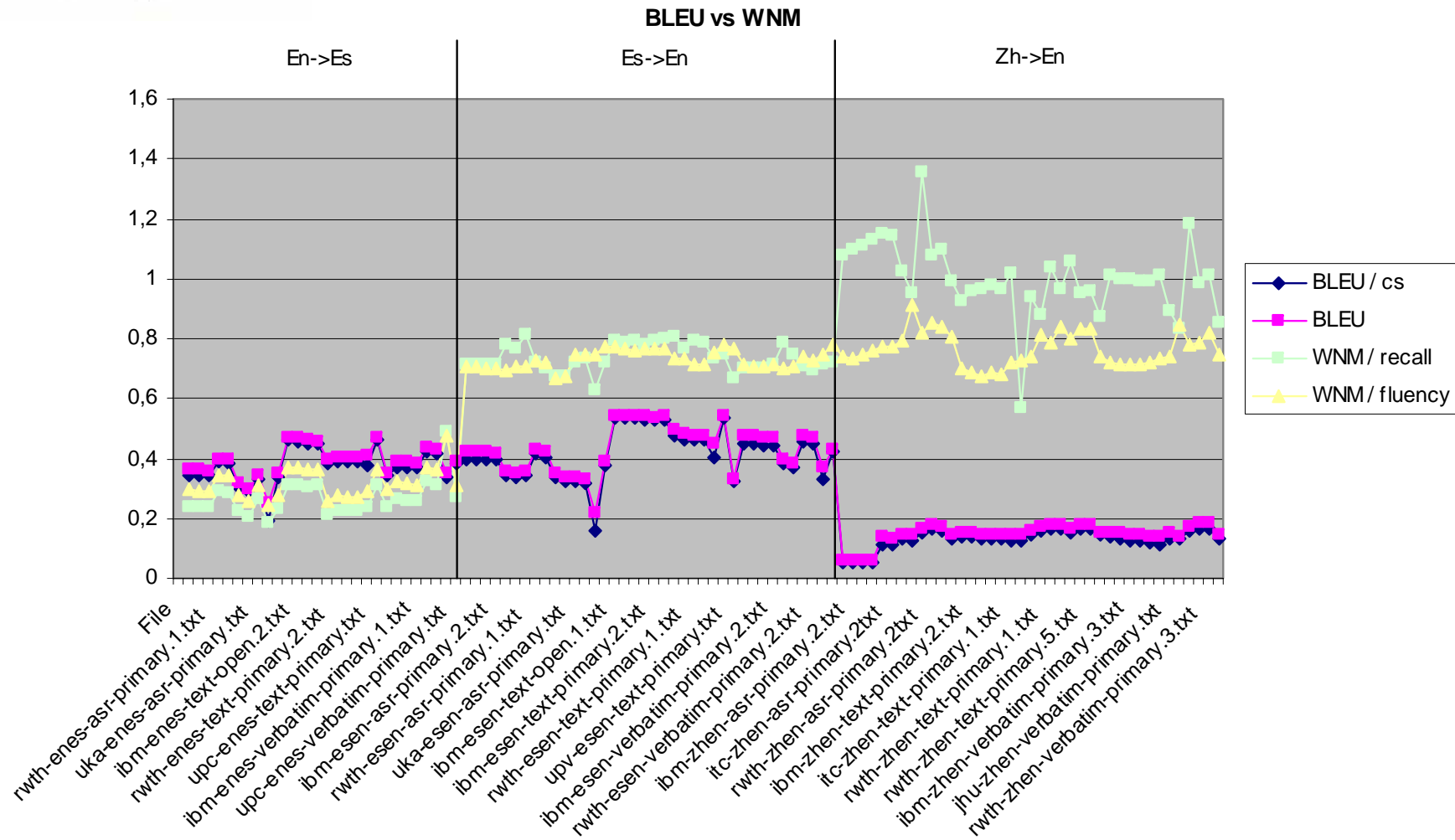


SLT Evaluation

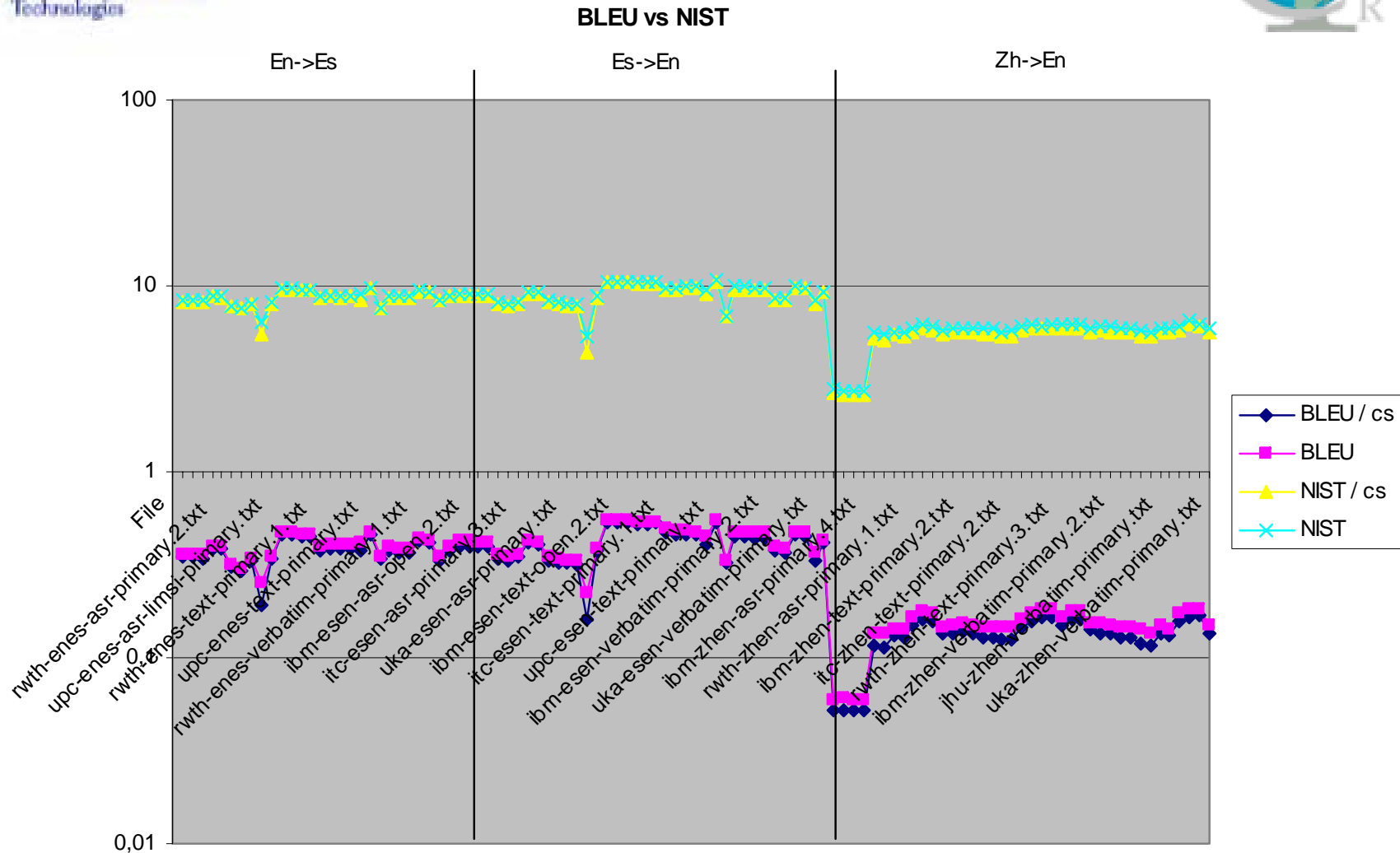


- For each task and condition, more than 1 submission allowed per participant
- Scoring
 - Metrics: BLEU/NIST v11, mWER, mPER, mCER
 - Scoring tools developed by ELDA (except BLEU/NSIT)
 - Other metrics will be available later on (X-Score, D-Score)
- Human evaluation? To be set up.

SLT Results – BLEU vs WNM



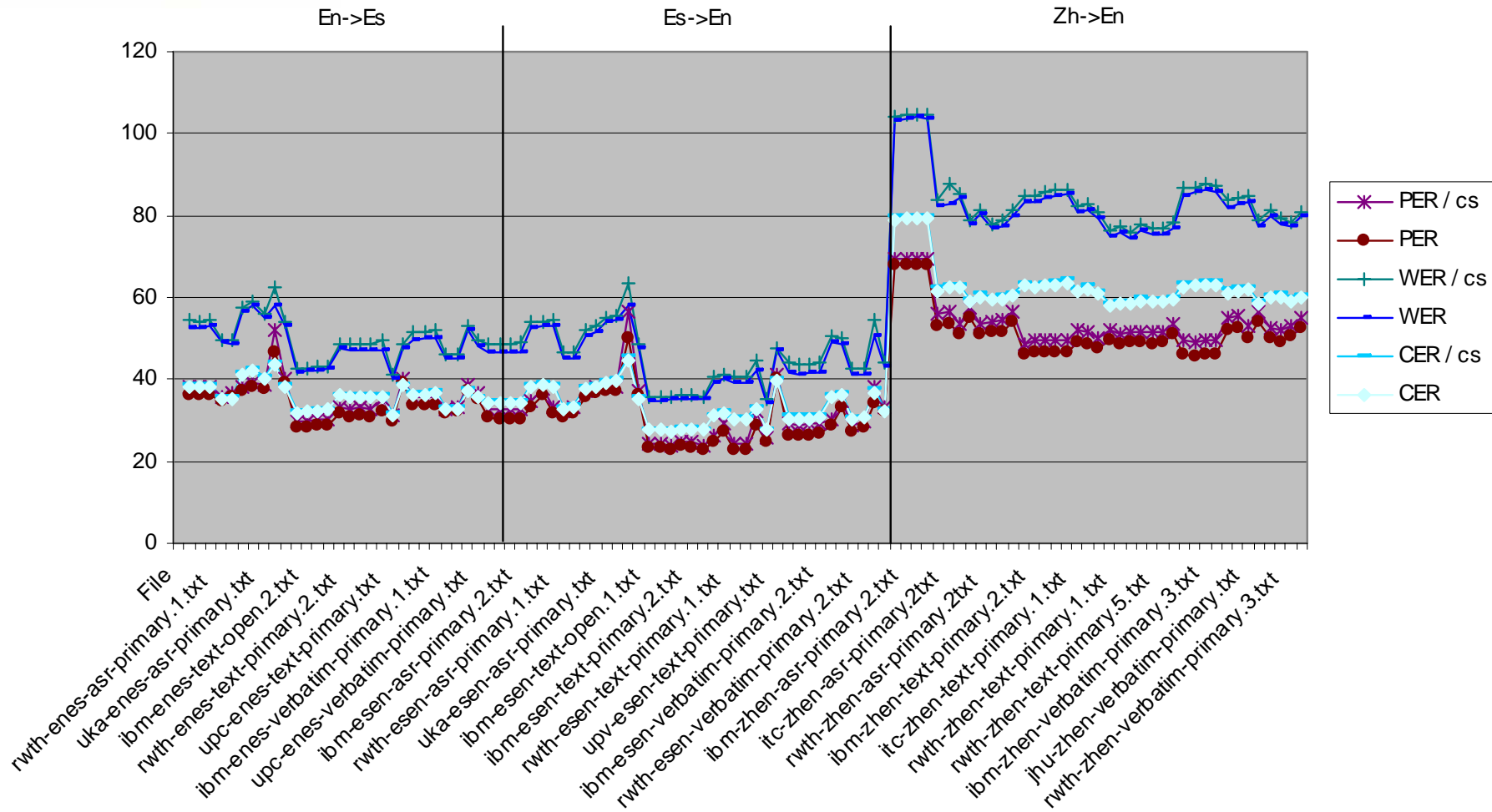
SLT Results – BLEU vs NIST



SLT Results – WER, PER, CER



WER vs PER vs CER



Conclusions (1)



- This is a good start for a European evaluation platform
- Evaluation packages to be made available
 - ASR
 - EPPS train, dev and test set
 - Scoring packages for En, Es and Zh
 - SLT
 - EPPS English-to-Spanish: EPPS training, development and test (incl. ref. translations) sets
 - EPPS Spanish-to-English: EPPS training, development and test (incl. ref. translations) sets
 - VOA Mandarin-to-English: VOA development and test sets (incl. resegmented transcripts and ref. translations)

Conclusions (2)



- Quality of data was not good enough
- Validation occurred too late
- Suggestions
 - Re-run SLT evaluation on corrected data
 - Guidelines should be more precisely defined (ambiguous interpretations)
 - Data production should be better scheduled and start earlier
 - Data validation should be scheduled within the production cycle and before full use



TTS Evaluation



1st Evaluation Campaign on Speech Synthesis

Antonio Bonafonte (UPC) and Marie Neige García (ELDA)

October 14, 2005

UPC

TASKS



1. Evaluation of systems (UPC: Spanish)
2. Evaluation of modules (M1, M2, M3). Well defined modules allow to participate groups which are experts on some crucial parts of speech synthesis.
 - M1: Text analysis (SIEMENS: English; NOKIA: Mandarin)
 - M2: Prosody (UPC: Spanish; NOKIA: Mandarin)
 - M3: Speech Generation
3. Specific evaluation of research activities
 - Evaluation of voice conversion (SIEMENS: Spanish)
 - Evaluation of expressive speech in ST

CORPORA

English	50% written data (newspapers) 50% EPPS data : transcriptions of European parliament
Mandarin	863 program data (National High-tech program 863 TTS evaluation in 2003)
Spanish	EPPS data : transcriptions of European parliament

Corpora used for the evaluation

EVALUATION TESTS

Module 1: Text analysis

- Test M1.1 Evaluation of text normalization and end of sentence detection
- Test M1.2 Evaluation of word segmentation (Mandarin)
- Test M1.3 Evaluation of POS tagger
- Test M1.4 Evaluation of Pronunciation

Module 2: Prosody

- Test M2.1 Evaluation of prosody (using segmental information, resynthesis)
- Test M2.2 Judgment test using delexicalized utterances
- Test M2.3 Functional test using delexicalized utterances (identify written sentences which the produced delexicalized prosody)

Module 3: Acoustic generation

- Test M3.1 Intelligibility (functional test)
- Test M3.2 Naturalness

Table 1: Evaluation Test for Module Evaluation

EVALUATION TESTS

System evaluation

Test S System evaluation (based on ITU P.85), MOS

Table 2: Evaluation of TTS component

Voice conversion

Test VC.1 Voice conversion *removing* prosody effect

Test VC.2 Voice conversion *including* prosody

Expressive speech

Test E Judgement test about speech expressivity

Table 3: Evaluation Tests for Specific Research Activities

- Corpus: 12 paragraphs from EPPS.
- *Systems:*
 - System 1: AT&T voice (female)
 - System 2: Festival voice, not tuned to the application (female)
 - System 3: UPC baseline voice (female)
 - System 4: UPC baseline voice (male)
- Human evaluation (17 subjects; 12 evaluations/system)

System	Quality	Effort	Pronun.	Compre.	Artic.	S. rate	Natural	Ease list.	Pleasant
1	4.23	4.39	4.01	4.58	4.15	4.20	3.16	3.53	3.64
2	2.60	2.96	2.72	3.22	3.60	3.32	2.29	2.09	2.90
3	3.84	4.14	4.02	4.35	3.94	4.38	3.07	3.19	3.42
4	4.07	4.28	4.29	4.52	4.25	4.38	3.47	3.48	3.89

Preliminar results computed but need to be confirmed


Categories	Abbreviations, letter-sequences, digit-sequences, cardinal, ordinal, time, date, money, punctuation, etc.
Corpus	100 samples/category found in data.
Reference	<i>Manual check</i>
Metric	WER

Table 4: Word Normalization

Corpus	500 sentences (EPSS + written data)
Reference	Manual segmentation.
Metric	Error rate

Table 5: End of Sentence

TEXT PROCESSING ENGLISH / SIEMENS

Corpus	10K words (EPSS + written data)
Reference	POS tagged manually
Metric	POS Error rate

Table 6: POS tagger

Corpus	2K words (50% common words; 25% geographic locations; 25% names)
Reference	Manual phonetisations, including alternatives
Metric	WER and phoneme error rate (Ph-ER)
Remark	Need for <i>normalization</i>

Table 7: Grapheme-Phoneme conversion



TEXT PROCESSING MANDARIN NOKIA



Similar to English but:

- Text normalisation (digits, time, date, measures)
- Word segmentation

and Data based on 836 program data: not ready

PROSODY SPANISH UPC

- Corpus: 12 paragraphs from EPPS, distributed over *melodic* domains.
- *Systems*: REF-M (Male Speaker), REF-F, UPC-M, UPC-F
- Human evaluation (17 subjects; 40 evaluations/system)

System/Test	Resynthesis (1 - 5)	delex. judgement (1 - 5)	delex. funct (0 - 100)
Ref-F	3.0	3.7	87%
Ref-M	4.0	4.0	76%
UPC-F	2.2	3.2	65%
UPC-M	2.4	2.8	55%

Table 8: Evaluation of Prosody (Spanish)



PROSODY MANDARIN NOKIA



Similar to English but:

- Only the delexicalized tests
- DTD (module specification) need to be adapted for Mandarin (e.g.: duration of syllable, not phoneme)

and Data based on 836 program data: not ready

INTRALINGUAL VOICE CONVERSION SPANISH SIEMENS



- 4 transformations: F1 (female 1) \rightarrow F2, F1 \rightarrow M1, M1 \rightarrow F1, M1 \rightarrow M2
- *Systems*: VTLN (linear transformation + VTLN), ResPred (linear transformation + residual prediction)
- Human evaluation (17 subjects):
 - Voice conversion score (VCS): 0: identical; 1: completely different.
 - Speech quality (MOS)

System/Test	VCS	MOS
Human Voice	1.00	4.6
VTLN	0.79	3.2
ResPred	0.57	1.7

Table 9: Evaluation of Prosody (Spanish)

NEXT ACTIVITIES



- Task 3.1 Baseline systems for research and evaluation of speech synthesis
- Task 3.2 Integration of speech synthesis in the speech-to-speech system
- Task 3.3 Voice conversion, manipulation and compression
- Task 3.4 Prosody modelling and expressive speech

New topics

- Integration (end-to-end systems)
- Cross-lingual voice conversion
- Expressive speech on SLT



In conclusion... progress in TC-STAR



FIRST EVALUATION CAMPAIGN ON ASR AND SLT

February- APRIL 2005

- Technology and preliminary results presented in Trento Workshop
- Five SLT research systems for EPPS task
- SLT evaluation with **real-life data**
- ASR on EPPS (English): **10% WER**
- Joint ASR + MT : **35% PER**

FIRST EVALUATION CAMPAIGN ON SPEECH SYNTHESIS:

June- September 2005.

- Technology and preliminar results presented as a satellite event of the conference:
Speech Analysis, Synthesis and Recognition:
Applications of Phonetics, Krakow, 23 September 2005,
www.staff.amu.edu.pl/~fonetyka/program.html



Progress in TC-STAR



Standardization activity

- UIMA architecture has been adopted
- First prototypes by the end of the year

Next year's events



[TC-STAR Open Lab on Speech Translation,
Trento 29-31 March 2006](#)

Satellite event of EUROPEAN ACL

[TC-STAR Second Evaluation Campaign](#)

Open Workshop Barcelona Spain 19-21 June

***YOU ARE ALL INVITED TO JOIN THE
EVALUATION AND/OR SUBMIT PAPERS !!***

WWW.TC-STAR.ORG

In conclusion...



- Europe with his 20 languages is a *unique* opportunity
- Let's focus in the next 10 years on this problem as a priority
- There are tons of real data produced everyday
- Let's define clear measurable objectives
- Let's measure the progress based on evaluation (**an European infrastructure is needed**)





The first evaluation took place in March 2005 for ASR and SLT and in September 2005 for TTS.

TC-STAR welcomes outside participants in its 2nd evaluation of January-February 2006. This participation is free of charge.

The TC-STAR 2006 evaluation campaign will consider:

- SLT in the following directions
 - Chinese-to-English (Broadcast News)
 - Spanish-to-English (European Parliament plenary speeches)
 - English-to-Spanish (European Parliament plenary speeches)
- ASR in the following languages
 - English (European Parliament plenary speeches)
 - Spanish (European Parliament plenary speeches)
 - Mandarin Chinese (Broadcast News)
- TTS in Chinese, English, and Spanish under the following conditions:
 - Complete system: participants use their own training data
 - Voice conversion intralingual and crosslingual, expressive speech: data provided by TC-STAR
 - Component evaluation



Information Society
Technologies

TC-STAR EVALUATION CAMPAIGN 2006

Barcelona Workshop june 19-21



- For ASR and SLT, training data will be made available by the TC-STAR project for English and Spanish and can be purchased at LDC for Chinese. Development data will be provided by the TC-STAR project. Legal issues regarding the data will be detailed in the 2nd Call For Participation.
- All participants will be given the opportunity to present and discuss their results in the TC-STAR evaluation **workshop in Barcelona in 19-21 June 2006 at UPC**

Tentative schedule:

- Registration: November 2005 (early expression of interest is welcome)
ASR evaluation: from begin February to mid February 2006
SLT evaluation: from mid February to end February 2006
TTS evaluation: from begin February to end of February 2006
- Release: April 2006
Submission of papers: May 2006
Workshop: June 19-21 2006
- Contact: Djamel Mostefa (ELDA)
e-mail: mostefa@elda.org
tel. +33 1 43 13 33 33

TC-STAR Open Lab on Speech Translation: Trento

29-31 march 2006 – Before EACL 2006

First Call For Participation

- This event is sponsored by the European Integrated Project TC-STAR (Technologies and Corpora for Speech-to-speech Translation Research).
- It aims to expand outside the TC-STAR research community and to work in the areas of Automatic Speech Recognition (ASR) and Spoken Language Translation (SLT).
- Students and researchers in the field of human language technology are invited to contribute to the following topics proposed by the organizers:
 - * Integration of ASR and SLT
 - * System combination in ASR and SLT
 - * Morphology and Syntax in SLT
 - * Error analysis in SMT

TC-STAR Open Lab on Speech Translation: Trento 29-31 march 2006 - Before EACL 2006

- Some months before the meeting in Trento, shared tasks will be defined and language resources and tools for them will be made available to registered participants.
- The considered application domain will be the translation of European Parliament speeches from Spanish to English, and vice versa. For both tasks, word graphs and n-best lists generated by different ASR and SLT systems will be provided.
- Training and testing collections to develop and evaluate a SLT system will be distributed, too. Participants will be given the opportunity to present and discuss their results at the meeting in Trento and to attend tutorials held by experts in the field.
- A limited number of grants will be made available to students and junior researchers to cover lodging and food expenses.
- Information about the Open Lab will be published at: <http://www.tc-star.org>.
- **Organizers:** Marcello Federico, ITC-irst, Trento
Ralf Schlüter, RWTH, Aachen
- **Contact:** openlab2006@tc-star.org