Report on the IWSLT 2013 Evaluation Campaign

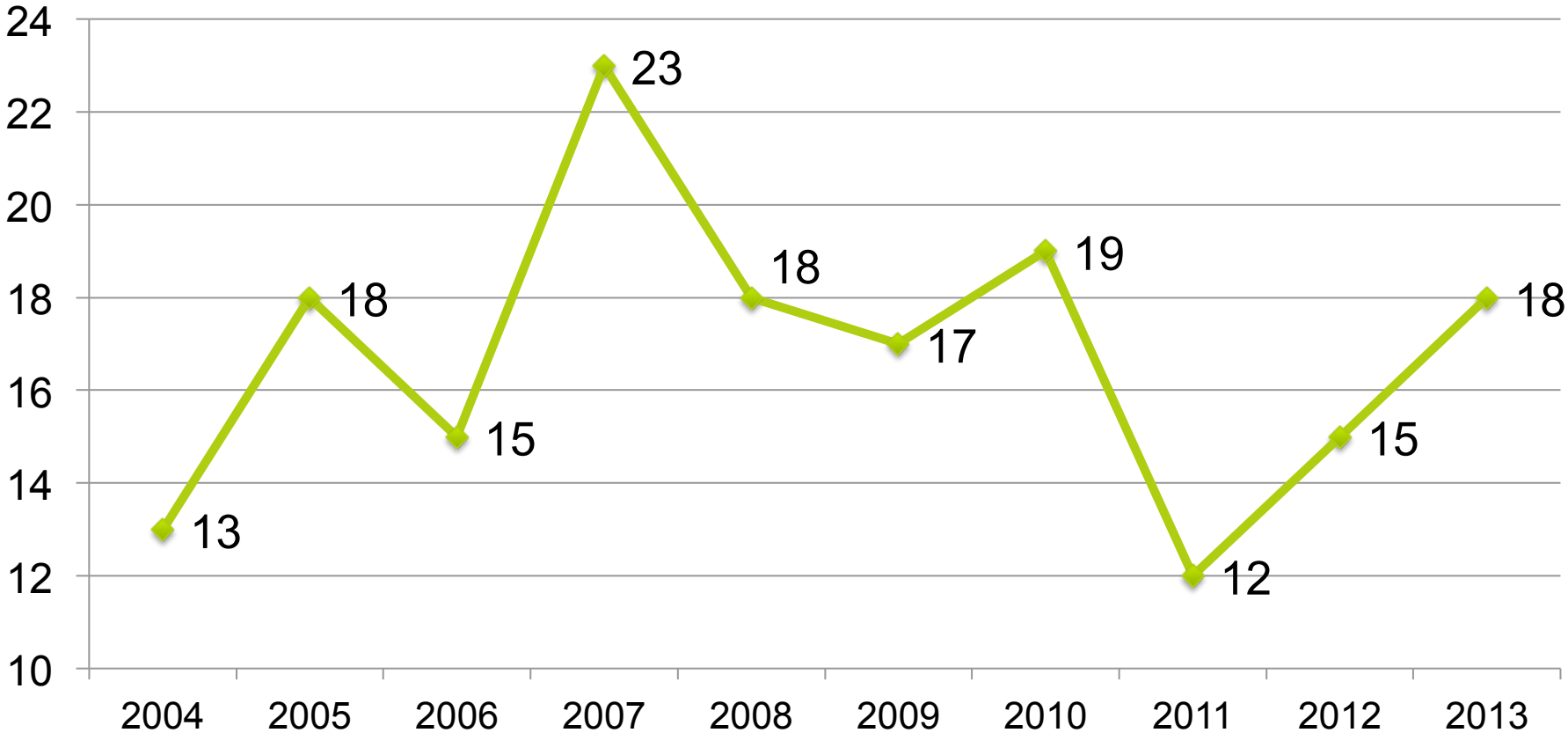
Mauro Cettolo, FBK-irst, Italy
Jan Nieheus, KIT, Germany
Sebastian Stueker, KIT, Germany
Luisa Bentivogli, CELCT, Italy
Marcello Federico, FBK-irst, Italy

IWSLT, Heidelberg, 5-6 December 2013

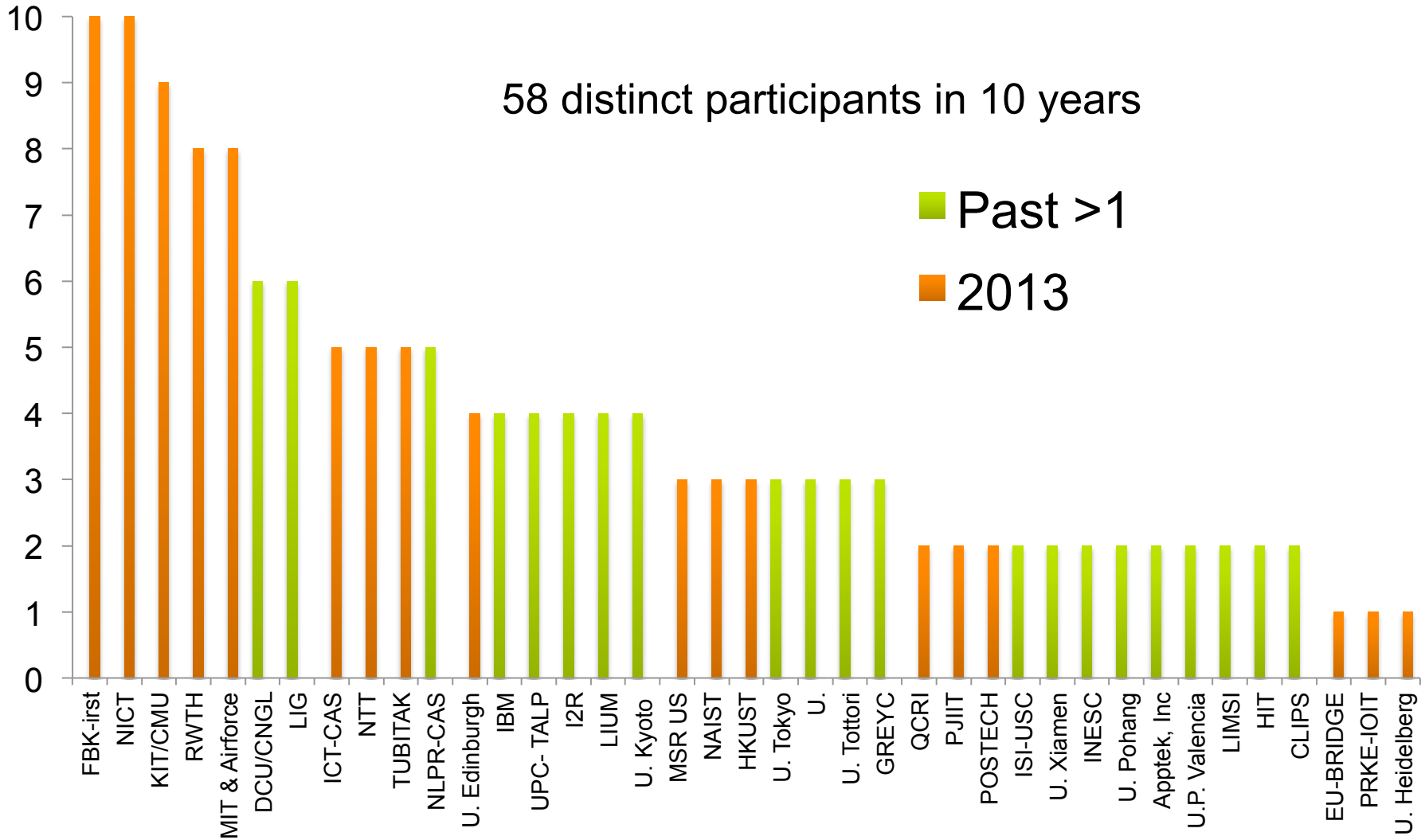
Outline

- IWSLT review
- TED Talks
- Tracks
- Automatic evaluation
- Human evaluation
- Future plans

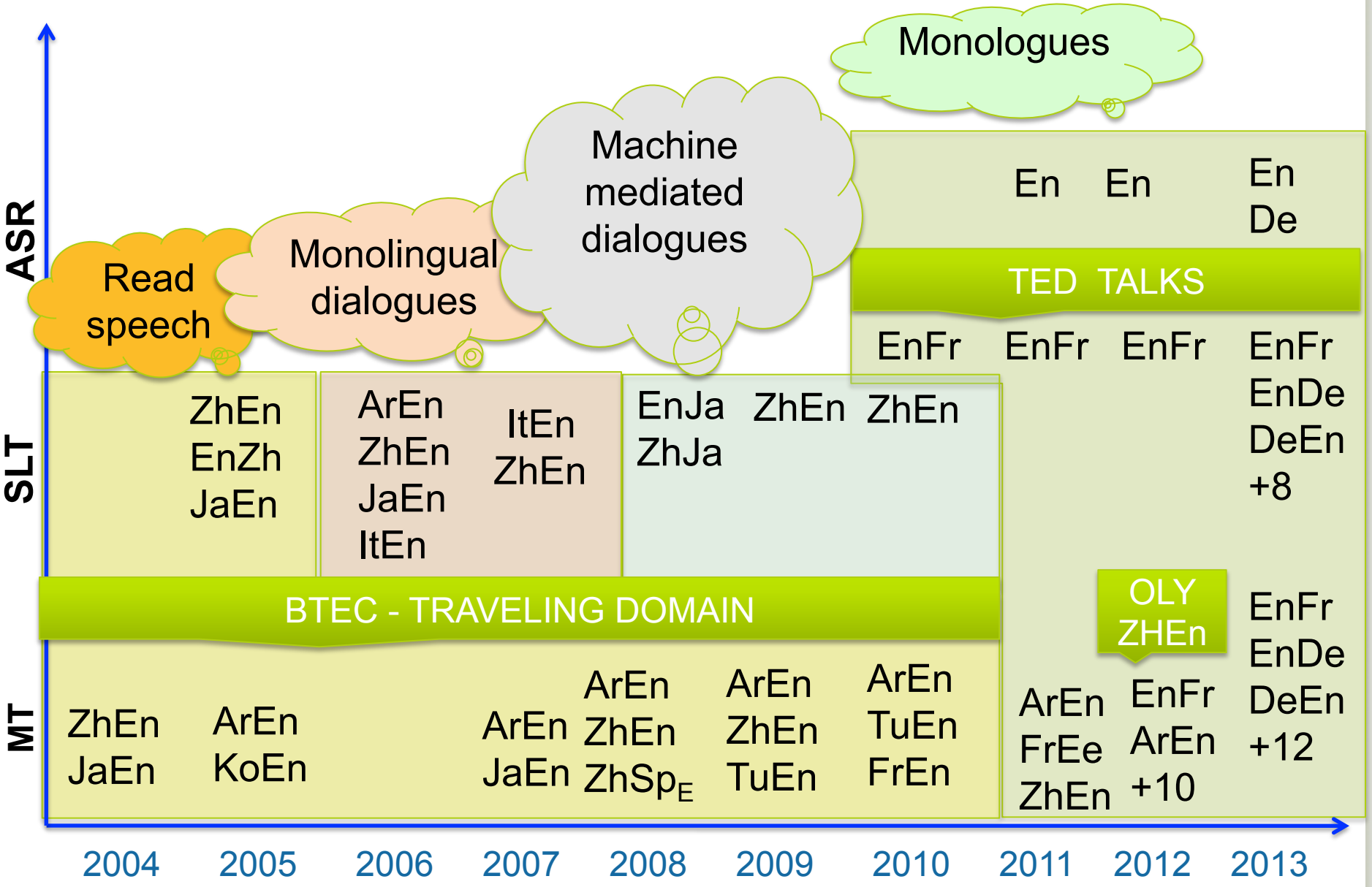
IWSLT Evaluation: record of participants



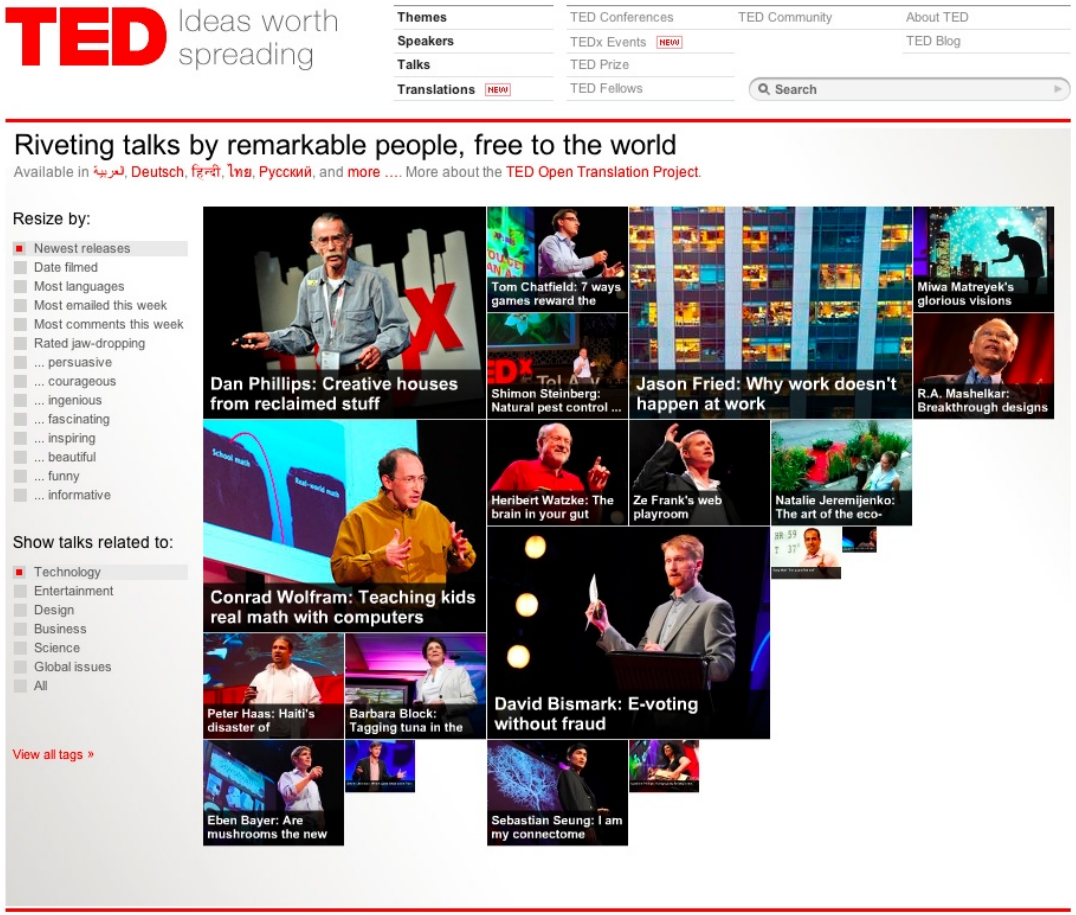
IWSLT Evaluation: record of participants



IWSLT: tasks and languages

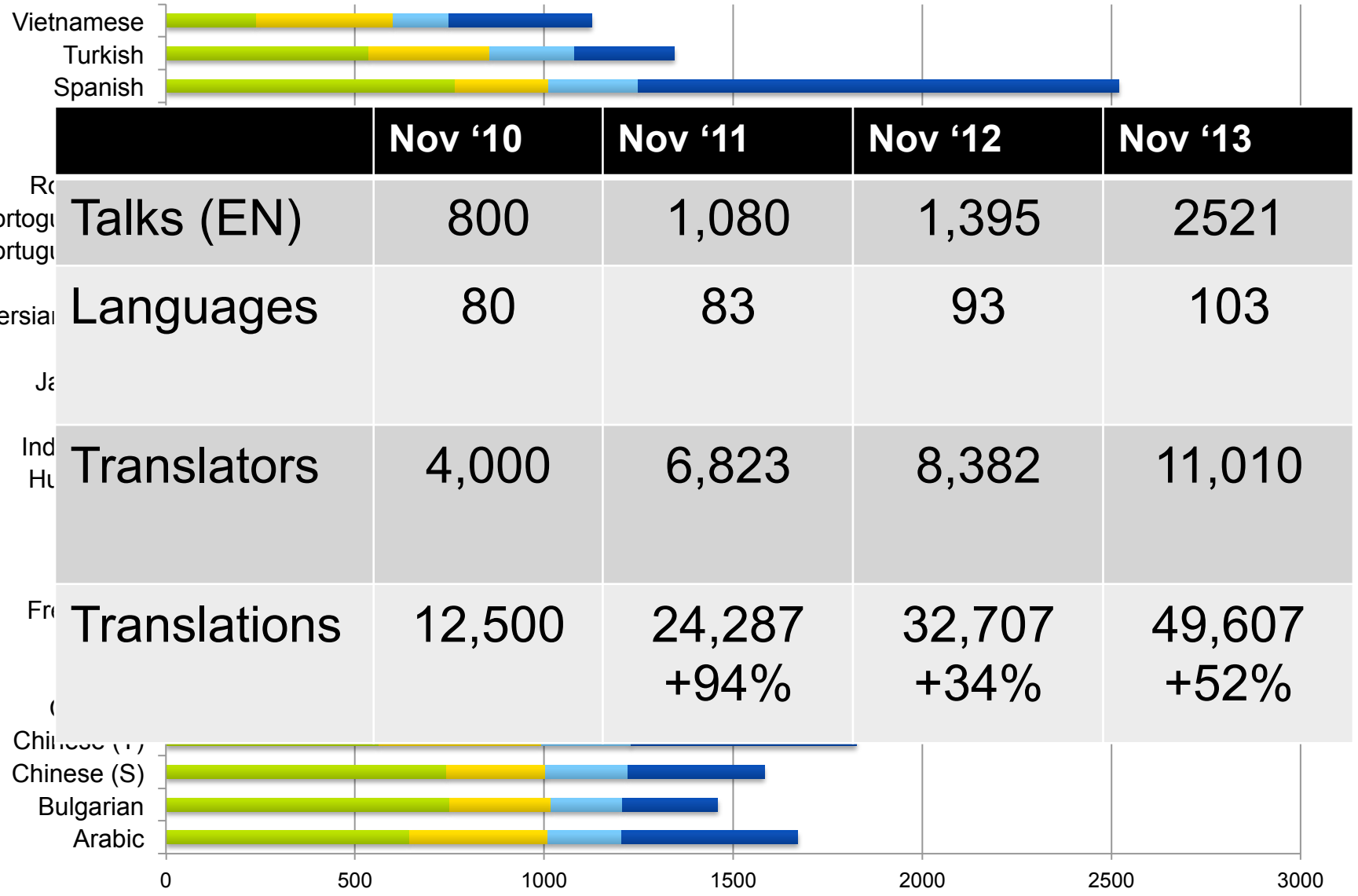


TED Talks



- TED LLC is non-profit
- Two annual events
- Short talks
- Variety of topics
- Website with:
  - Videos
  - Transcripts
  - Translations
- CC License

TED Talks Translations (from English)



	Nov ‘10	Nov ‘11	Nov ‘12	Nov ‘13
Talks (EN)	800	1,080	1,395	2521
Languages	80	83	93	103
Translators	4,000	6,823	8,382	11,010
Translations	12,500	24,287 +94%	32,707 +34%	49,607 +52%

Human task: subtitling and translating

- ✓ segment audio
- ✓ transcribe and annotate
- ✓ split into captions
- ✓ translate captions

Challenges in TED Task

- **Language modelling**
  - Limited in-domain training data
  - Variability of topics and styles
- **Acoustic modelling**
  - Speaker: accent, fluency, speaking rate, style, , ...
  - Noise: mumble, applauses, laughs, music, ...
- **Translation modelling**
  - Distant and under-resourced languages
  - Morphologically rich languages
- **Speech Translation**
  - From spontaneous speech to polished text
  - Detection and removal of non-speech events
  - Subtitling and translating in real-time

Challenges for 2011

- **Language modelling**
  - Limited in-domain training data
  - Variability of topics and styles
- **Acoustic modelling**
  - Speaker: accent, fluency, speaking rate, style, , ...
  - Noise: mumble, applauses, laughs, music, ...
- **Translation modelling**
  - Distant and under-resourced languages
  - Morphologically rich languages
- **Speech Translation**
  - From spontaneous speech to polished text
  - Detection and removal of non-speech events
  - Subtitling and translating in real-time

Challenges for 2012

- **Language modelling**
  - Limited in-domain training data
  - Variability of topics and styles
- **Acoustic modelling**
  - Speaker: accent, fluency, speaking rate, style, , ...
  - Noise: mumble, applauses, laughs, music, ...
- **Translation modelling**
  - Distant and **under-resourced** languages
  - **Morphologically rich languages**
- **Speech Translation**
  - From spontaneous speech to polished text
  - Detection and removal of non-speech events
  - Subtitling and translating in real-time

Challenges for 2013

- **Language modelling**
  - Limited in-domain training data
  - Variability of topics and styles
- **Acoustic modelling**
  - Speaker: accent, fluency, speaking rate, style, , ...
  - **Noise: mumble, applauses, laughs, music, ...**
- **Translation modelling**
  - Distant and under-resourced languages
  - Morphologically rich languages
- **Speech Translation**
  - From spontaneous speech to polished text
  - **Detection and removal of non-speech events**
  - Subtitling and translating in real-time

2013 Tracks

- **Automatic Speech Recognition (ASR)**
  - Transcription of talks from audio to text
  - English, **German**
- **Spoken Language Translation (SLT)**
  - Translation of talks from audio (or ASR output) to text
  - English-French, **English-German, German-English**
  - English-Arabic, English-Chinese **unofficial pairs**
- **Machine Translation (MT)**
  - Translation of talks from text to text
  - English-French, English-German, German-English
  - + X-English and English-X **unofficial pairs**

X= Arabic, Spanish, Portuguese (B), Italian, Chinese,
Polish, Persian, Slovenian, Turkish, Dutch, Romanian, Russian

Specifications

Conditions	ASR	SLT	MT
Input: Pre-segmented	yes	yes	yes
Input: Cased & Punctuated		no	yes
Output: Cased & Punctuated	no	yes	yes
Automatic evaluation (1)	yes	yes	yes
Human evaluation (En-FR)			yes

Metrics	ASR	SLT	MT
WER	✔	✔	✔
BLEU		✔	✔
TER		✔	✔

(1) Prepared non trivial reference baselines for all MT directions.

Participants

NTT-NAIST	NTT Communication Science Labs, Japan & NAIST[11]
KIT	Karlsruhe Institute of Technology, Germany [12, 13]
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen, Germany [14, 15]
EU-BRIDGE	RWTH& UEDIN& KIT& FBK[16]
HDU	Dept. of Computational Linguistics, Heidelberg University, Germany [17]
UEDIN	University of Edinburgh, UK [18, 19, 20]
FBK	Fondazione Bruno Kessler, Italy [21, 22]
PRKE-IOIT	Inst. of Inform. and Techn., Vietnamese Academy of Science and Technology [23]
POSTECH	Pohang University of Science and Technology, Korea [24]
MITLL-AFRL	Mass. Institute of Technology/Air Force Research Lab., USA [25]
QCRI	Qatar Computing Research Institute, Qatar Foundation, Qatar [26]
MSR-FBK	Microsoft Corporation, USA, and FBK[27]
HKUST	Hong Kong University of Science and Technology, Hong Kong [28]
NICT	National Institute of Communications Technology, Japan [29, 30]
NAIST	Nara Institute of Science and Technology, Japan [31]
PJIIT	Polish-Japanese Institute of Information Technology, Poland [32]
CASIA	Institute of Automation, Chinese Academy of Sciences, China [33]
TUBITAK	TUBITAK - Center of Research for Advanced Technologies, Turkey

Results: ASR English

Run	TST13 WER%	TST12 WER%	TST11 WER%
NICT	13.5	8.6	7.9
KIT	14.4	9.6	9.3
MITLL-AFR	15.9	11.3	10.6
RWTH	16.0	11.3	10.2
NAIST	16.2	10.0	9.1
UEDIN	22.1	11.6	10.2
FBK	23.2	16.2	13.6
PRKE-IOIT	27.2	16.2	14.6

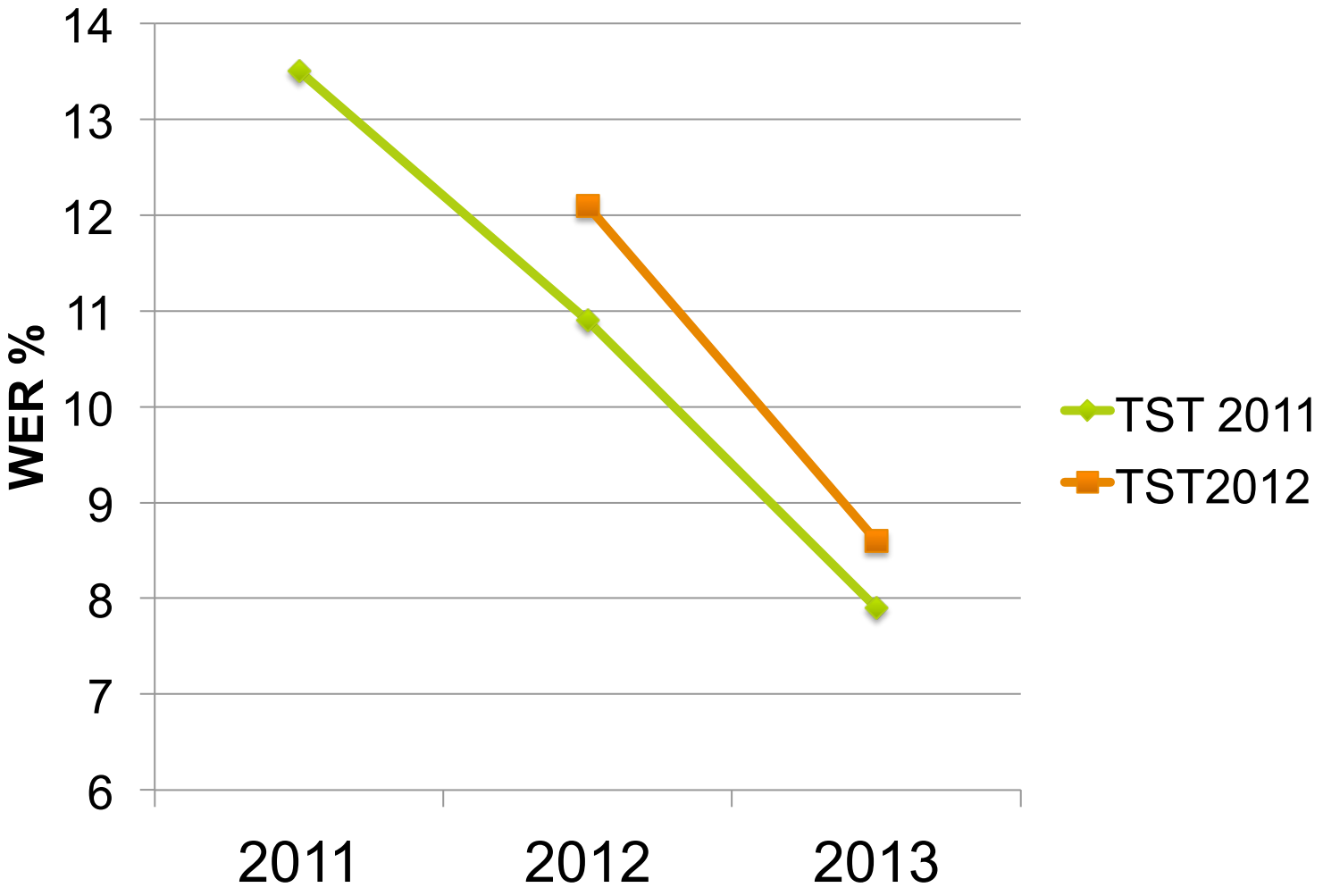
Results: ASR German

TED : ASR German ($\mathbf{ASR}_{DE}$)

System	WER	(# Errors)
RWTH	**25.2**	**(4,845)**
KIT	25.7	(4,932)
FBK	37.5	(7,199)
UEDIN	37.8	(7,250)

One outlier talk gives WER > 80%
Different training conditions among systems

Progress in ASR English (best systems)



Results: SLT

TED : SLT English-French (SLT_{EnFr})

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**26.81**	**55.08**	**27.53**	**54.06**
RWTH	25.62	57.21	26.41	56.09
UEDIN	22.45	61.34	23.30	60.06
MSR-FBK	22.42	63.69	23.72	62.20

Results: SLT

TED : SLT English-German (SLT_{EnDe})

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**18.05**	**64.46**	**18.66**	**63.22**
RWTH	17.27	66.33	17.88	65.09

TED : SLT German-English (SLT_{DeEn})

System	*Ref. with disfluencies*				Ref. without disfluencies			
	case sensitive		case insensitive		*case sensitive*		case insensitive	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
KIT	**19.34**	**62.27**	**19.80**	**61.34**	**19.54**	**62.74**	**20.01**	**61.80**
UEDIN	14.92	68.12	15.39	67.28	15.03	68.70	15.52	67.86

Results: MT

TED : MT English-French (MT_{EnFr})

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**38.86**	**42.96**	**39.74**	**42.02**
KIT	38.63	43.20	39.60	42.11
UEDIN	38.45	43.96	39.39	42.91
FBK	37.69	44.13	38.46	43.23
RWTH	37.67	44.00	38.49	43.04
PRKE-IOIT	37.59	45.07	38.39	44.15
MITLL-AFRL	37.05	45.36	38.27	44.10
BASELINE	31.94	48.59	32.56	47.75

Results: MT

TED : MT German-English (SLT_{DeEn})

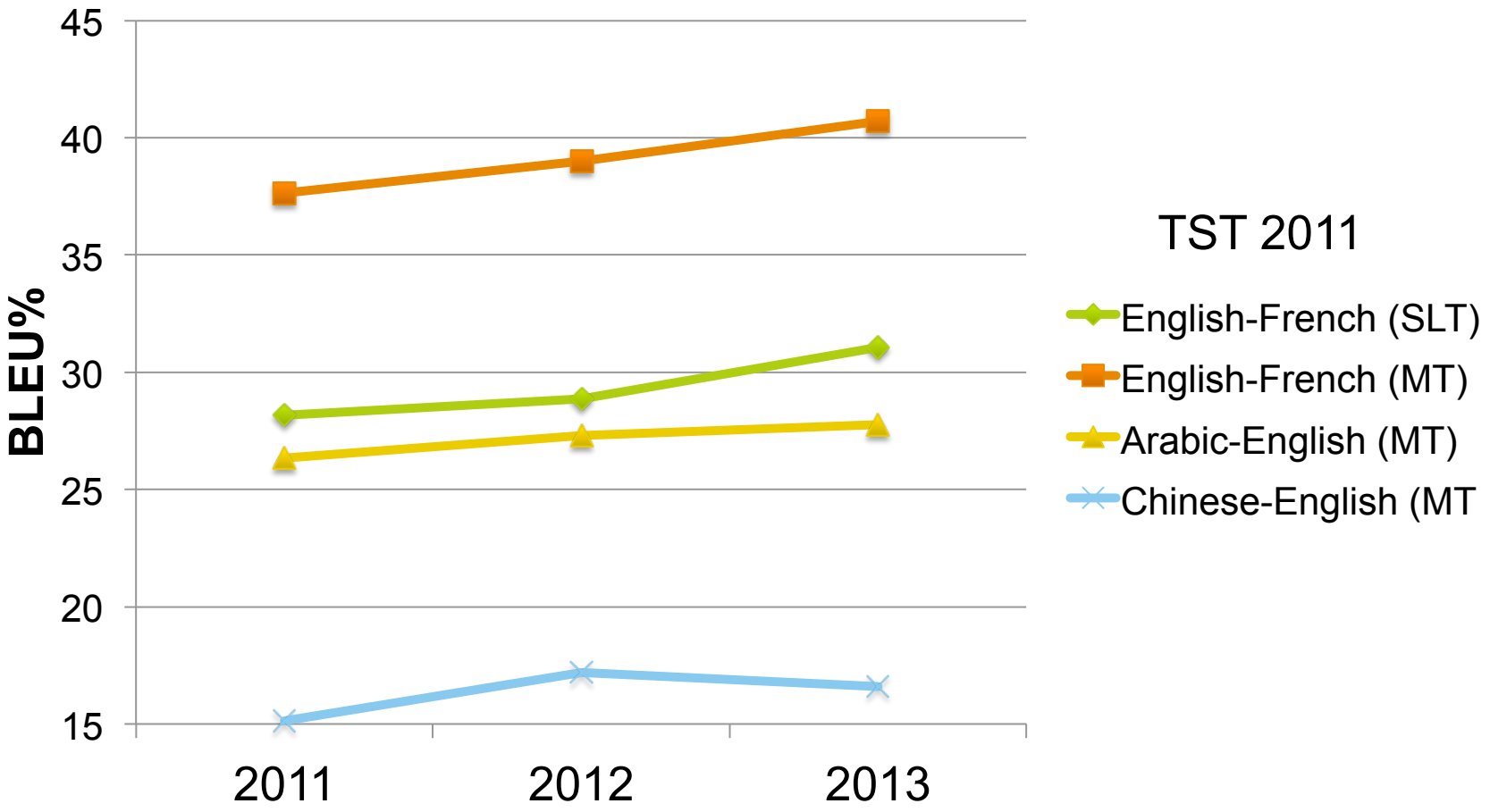
System	*Ref. with disfluencies*				Ref. without disfluencies			
	case sensitive		case insensitive		*case sensitive*		case insensitive	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
KIT	**26.48**	57.52	**27.11**	56.60	**26.57**	58.31	**27.16**	57.41
EU-BRIDGE	26.33	**56.70**	26.91	**55.78**	**26.57**	**57.29**	27.14	**56.38**
NTT-NAIST	25.69	60.96	26.29	60.06	25.83	60.75	26.45	59.82
UEDIN	25.54	59.99	26.12	59.07	25.35	60.98	25.87	60.08
RWTH	25.32	59.67	25.94	58.67	25.27	60.46	25.86	59.51
HDU	22.91	59.65	23.94	58.35	23.06	60.38	24.07	59.11
POSTECH	21.26	67.61	21.74	66.72	21.17	68.91	21.65	68.04
BASELINE	19.25	65.03	19.79	64.19	19.07	65.94	19.55	65.11

Results: MT

TED : MT English-German (MT_{EnDe})

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**25.71**	**54.46**	**26.47**	**53.34**
RWTH	24.74	55.52	25.41	54.42
NTT-NAIST	24.60	54.86	25.79	53.37
UEDIN	24.00	55.94	24.68	54.87
POSTECH	22.43	57.57	23.00	56.58
BASELINE	19.58	59.81	20.14	58.84

Progress in SLT and MT (best systems)



Human Evaluation (MT En-Fr)

New evaluation: ***Post-Editing + HTER***

- TED task as an interesting application scenario to test the utility of MT systems in a real subtitling task
- Additional reference translations
- Edits point to specific translation errors
- HTER correlates well with human judgments

- Performed on the 2012 progress test set *(tst2012)*
  - Human Evaluation (HE) Set
    - 580 sentences (~10,000 words)
    - initial 50% of 11 different talks

Evaluation Setup and Data Collection (1/2)

- *Bilingual* Post-Editing
  - professional translators were required to post-edit the MT output directly according to the source sentence
- Data preparation:
  - 7 systems p-edited by 7 professional translators
    - each translator must p-edit all the HE set sentences
    - each translator must p-edit each sentence only once
    - each MT system must be equally p-edited by all translators
  - MT outputs dispatched to translators both randomly and satisfying the uniform assignment constraints
- MateCat Project post-editing interface

Evaluation Setup and Data Collection (2/2)

- Collected data
  - 7 new references for each sentence in the HE set (1 *targeted* references for each MT system + 6 additional references)
- Post-editors variability:

PEditor	PE Effort	*std-dev*	Sys TER	*std-dev*
PE 1	24.93	17.74	40.27	20.32
PE 2	34.03	19.86	39.48	19.89
PE 3	42.60	22.47	40.61	20.19
PE 4	32.78	21.07	39.98	20.97
PE 5	19.51	15.55	40.82	20.95
PE 6	30.64	19.48	40.42	20.70
PE 7	34.60	23.92	39.39	20.62

- Evaluation: HTER calculated on **two** different evaluation settings
  - using the targeted reference only (*Tgt PEref*)
  - using all 7 references produced by the post-editors (*All PErefs*).

HTER Examples

SRC (2012.1417-18/19):
The only thing that matters is that you win or that you lose.
But why would you reconcile after a fight?

Only Targeted Reference

REF: La seule chose qui compte * ** est *** **de** **gagner** ou *** **de** perdre .
HYP: La seule chose qui compte **, c'** est **que vous gagnez** ou **que vous** perdez .

TER: 66.67

REF: Mais pourquoi voudriez-vous **vous réconcilier** après **vous être battu** ?
HYP: Mais pourquoi voudriez-vous **** **concilier** après **** **un** **combat** ?

TER: 50.00

All Post-Edited References

REF: La seule chose qui compte , c' est que vous gagnez ou que vous perdez .
HYP: La seule chose qui compte , c' est que vous gagnez ou que vous perdez .

TER: 0.00

REF: Mais pourquoi **se** **réconcilier** après un combat ?
HYP: Mais pourquoi **voudriez-vous concilier** après un combat ?

TER: 23.33

Evaluation Results

System Ranking	**HTER** ***HE Set all PErefs***	**HTER** HE Set Tgt PEref	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**18.67**	29.83	38.71	38.72
KIT	**20.01**	*29.64*	39.20	39.22
UEDIN	**20.69**	31.61	39.81	39.83
RWTH	**21.06**	31.64	*39.70*	39.95
FBK	**21.41**	32.29	40.38	40.56
MITLL-AFRL	**22.24**	32.31	41.37	41.47
PRKE-IOIT	**22.26**	*32.01*	41.81	41.52
Rank Corr.		.857	.964	1.00

Evaluation Results

System Ranking	HTER *HE Set all PErefs*	HTER HE Set Tgt PEref	TER HE Set ref	TER Test Set ref
EU-BRIDGE	**18.67**	29.83	38.71	38.72
KIT	**20.01**	*29.64*	39.20	39.22
UEDIN	**20.69**	31.61	39.81	39.83
RWTH	**21.06**	31.64	*39.70*	39.95
FBK	**21.41**	32.29	40.38	40.56
MITLL-AFRL	**22.24**	32.31	41.37	41.47
PRKE-IOIT	**22.26**	*32.01*	41.81	41.52
Rank Corr.		.857	.964	1.00

Statistical Significance (Approximate Randomization):

EU-BRIDGE vs. FBK: $p < 0.1$

vs. MITLL-AFRL: $p < 0.05$

vs. PRKE-IOIT: $p < 0.05$

Evaluation Results

System Ranking	**HTER** ***HE Set all PErefs***	**HTER** HE Set Tgt PEref	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**18.67**	29.83	38.71	38.72
KIT	**20.01**	*29.64*	39.20	39.22
UEDIN	**20.69**	31.61	39.81	39.83
RWTH	**21.06**	31.64	*39.70*	39.95
FBK	**21.41**	32.29	40.38	40.56
MITLL-AFRL	**22.24**	32.31	41.37	41.47
PRKE-IOIT	**22.26**	*32.01*	41.81	41.52
Rank Corr.		.857	.964	1.00

Spearman's Coefficient

Future plans

- Add more ASR languages
- Provide training data for acoustic modelling (BNs)
- Include more X-English SLT tasks (with audio)
  - Italian-English, French-English, … based on TEDx
- Ask participants to provide ASR real-time factor
- Release TST2013 and TST2012 with new references
- Release the baseline MT systems
- Measure progress on TST2011
- Continue with HE based on post-editing

Credits

- **Language resources**
  - TED LLC, USA (Talk data)
  - Workshop Machine Translation (Giga and news data)
  - DFKI, Germany (United Nations data)
- **Funding**
  - EUBRIDGE IST 287658
  - Concept for the Future, German Excellence Initiative

Questions?