Report on the IWSLT 2014 Evaluation Campaign

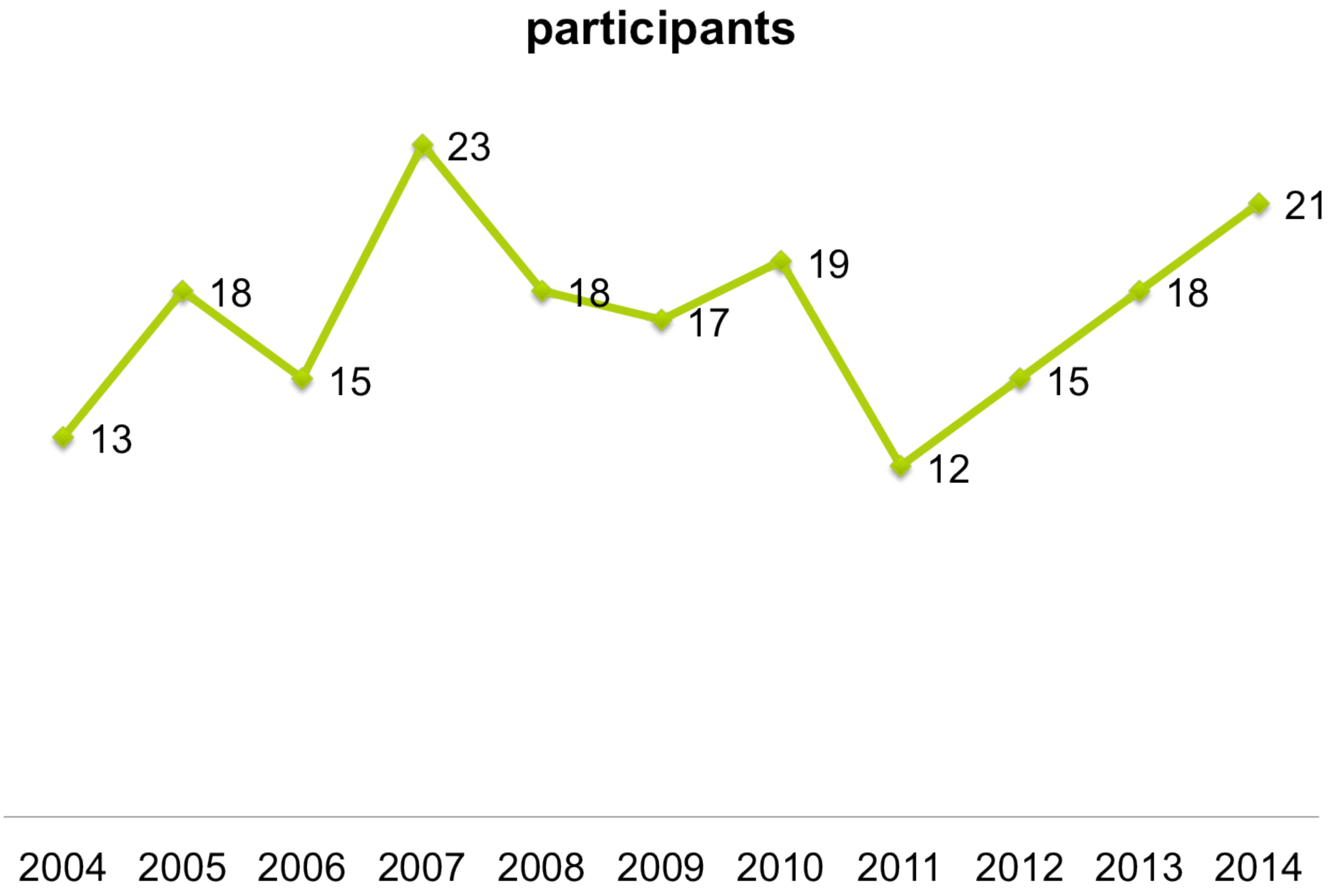
Mauro Cettolo, FBK-irst, Italy
Jan Niehues, KIT, Germany
Sebastian Stüker, KIT, Germany
Luisa Bentivogli, CELCT, Italy
Marcello Federico, FBK-irst, Italy

IWSLT, Lake Tahoe, 4-5 December 2014

Outline

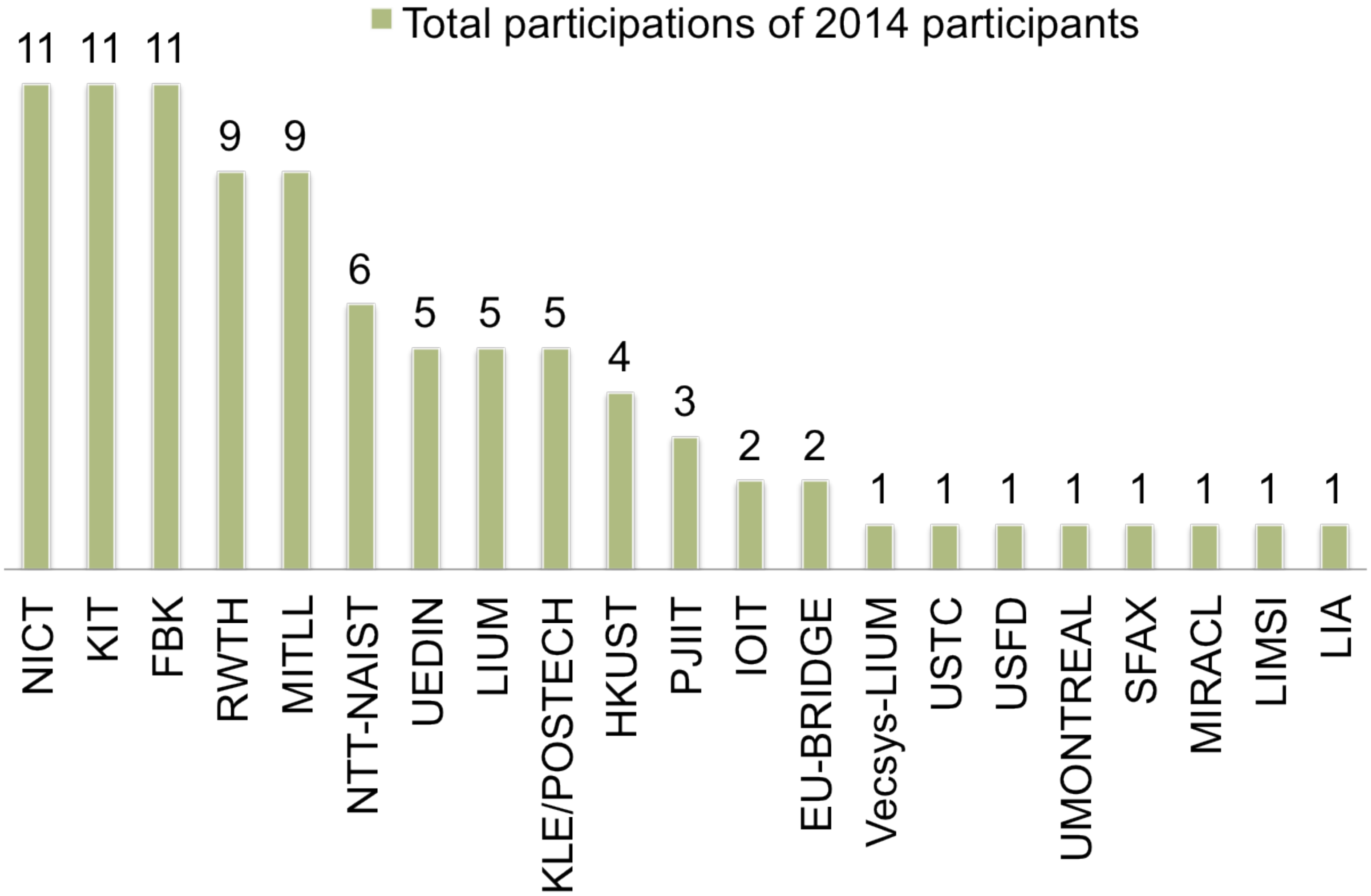
- IWSLT review
- TED Talks
- Tracks
- Automatic evaluation
- Human evaluation
- Future plans

IWSLT Evaluation: record of participants

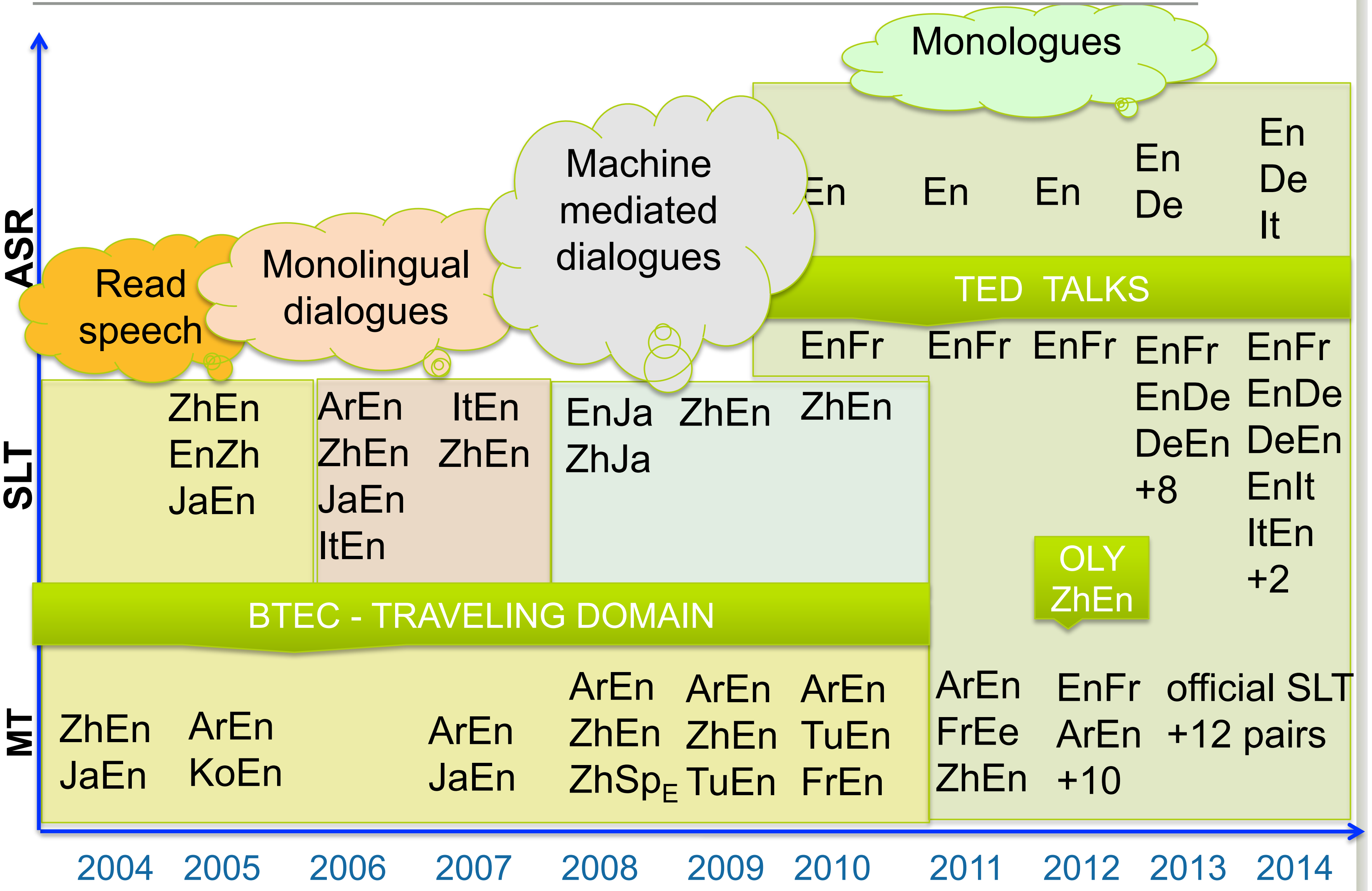


IWSLT Evaluation: record of participants

more than 60 distinct participants in 11 years



IWSLT: tasks and languages



ASR

SLT

MT

Read speech

Monolingual dialogues

Machine mediated dialogues

Monologues

En En En En De En De It

TED TALKS

EnFr EnFr EnFr EnFr EnDe DeEn +8 EnFr EnDe DeEn EnIt ItEn +2

OLY ZhEn

ZhEn EnZh JaEn

ArEn ZhEn JaEn ItEn

ItEn ZhEn

EnJa ZhJa

ZhEn

ZhEn

BTEC - TRAVELING DOMAIN

ZhEn JaEn

ArEn KoEn

ArEn JaEn

ArEn ZhEn $ZhSp_E$

ArEn ZhEn TuEn

ArEn TuEn FrEn

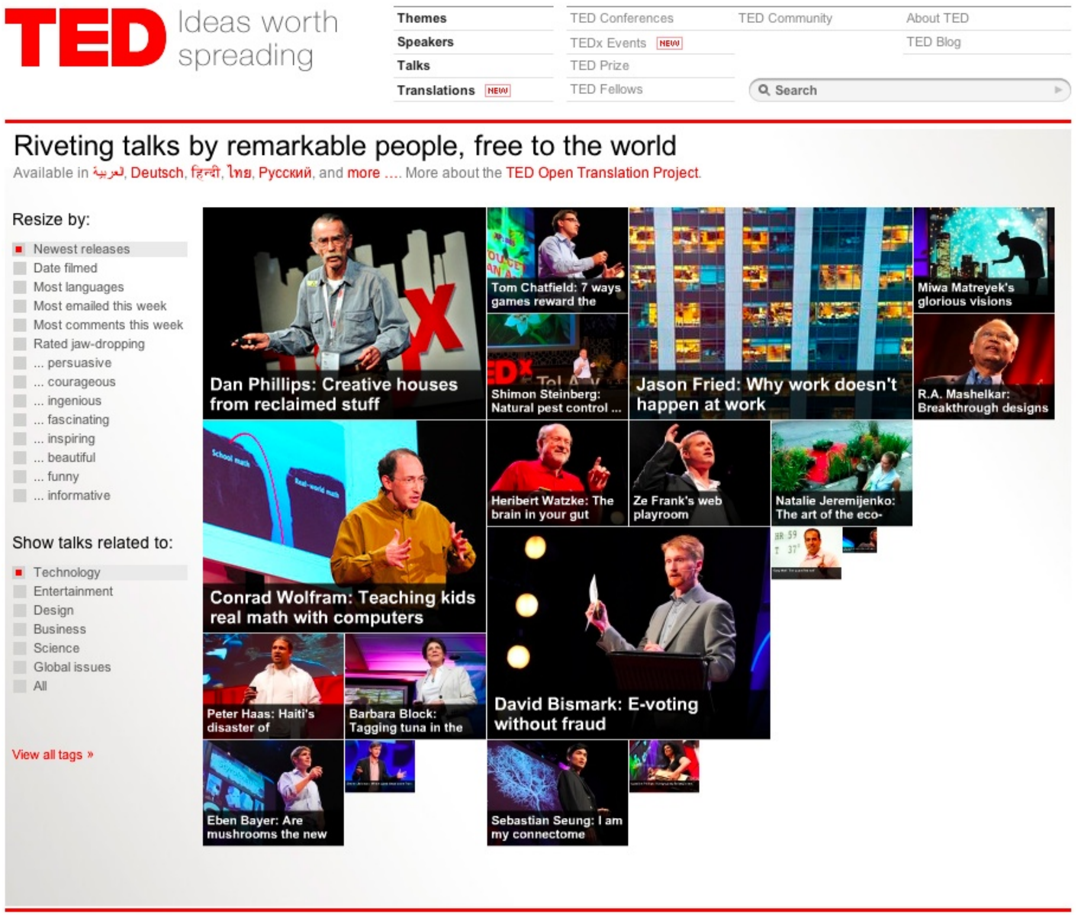
ArEn FrEe ZhEn

EnFr ArEn +10

official SLT +12 pairs

2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

TED Talks



Riveting talks by remarkable people, free to the world

Available in العربية, Deutsch, हिन्दी, ไทย, Русский, and more .... More about the TED Open Translation Project.

Resize by:

- ■ Newest releases
- Date filmed
- Most languages
- Most emailed this week
- Most comments this week
- Rated jaw-dropping
- ... persuasive
- ... courageous
- ... ingenious
- ... fascinating
- ... inspiring
- ... beautiful
- ... funny
- ... informative

Show talks related to:

- ■ Technology
- Entertainment
- Design
- Business
- Science
- Global issues
- All

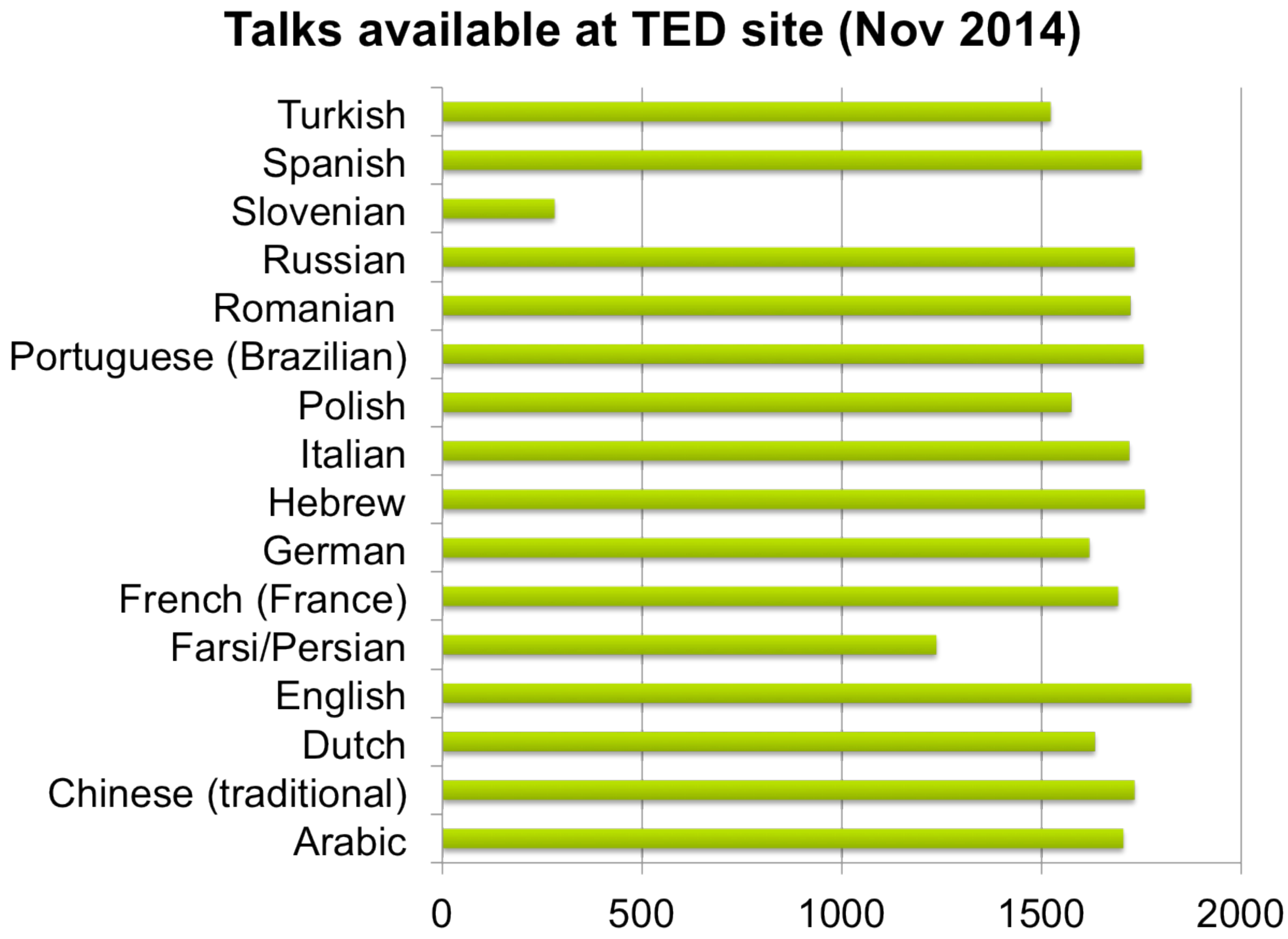
View all tags »

Dan Phillips: Creative houses from reclaimed stuff
Tom Chatfield: 7 ways games reward the
Jason Fried: Why work doesn't happen at work
Miwa Matreyek's glorious visions
Shimon Steinberg: Natural pest control ...
R.A. Mashelkar: Breakthrough designs
Conrad Wolfram: Teaching kids real math with computers
Heribert Watzke: The brain in your gut
Ze Frank's web playroom
Natalie Jeremijenko: The art of the eco-
David Bismark: E-voting without fraud
Peter Haas: Haiti's disaster of
Barbara Block: Tagging tuna in the
Eben Bayer: Are mushrooms the new
Sebastian Seung: I am my connectome

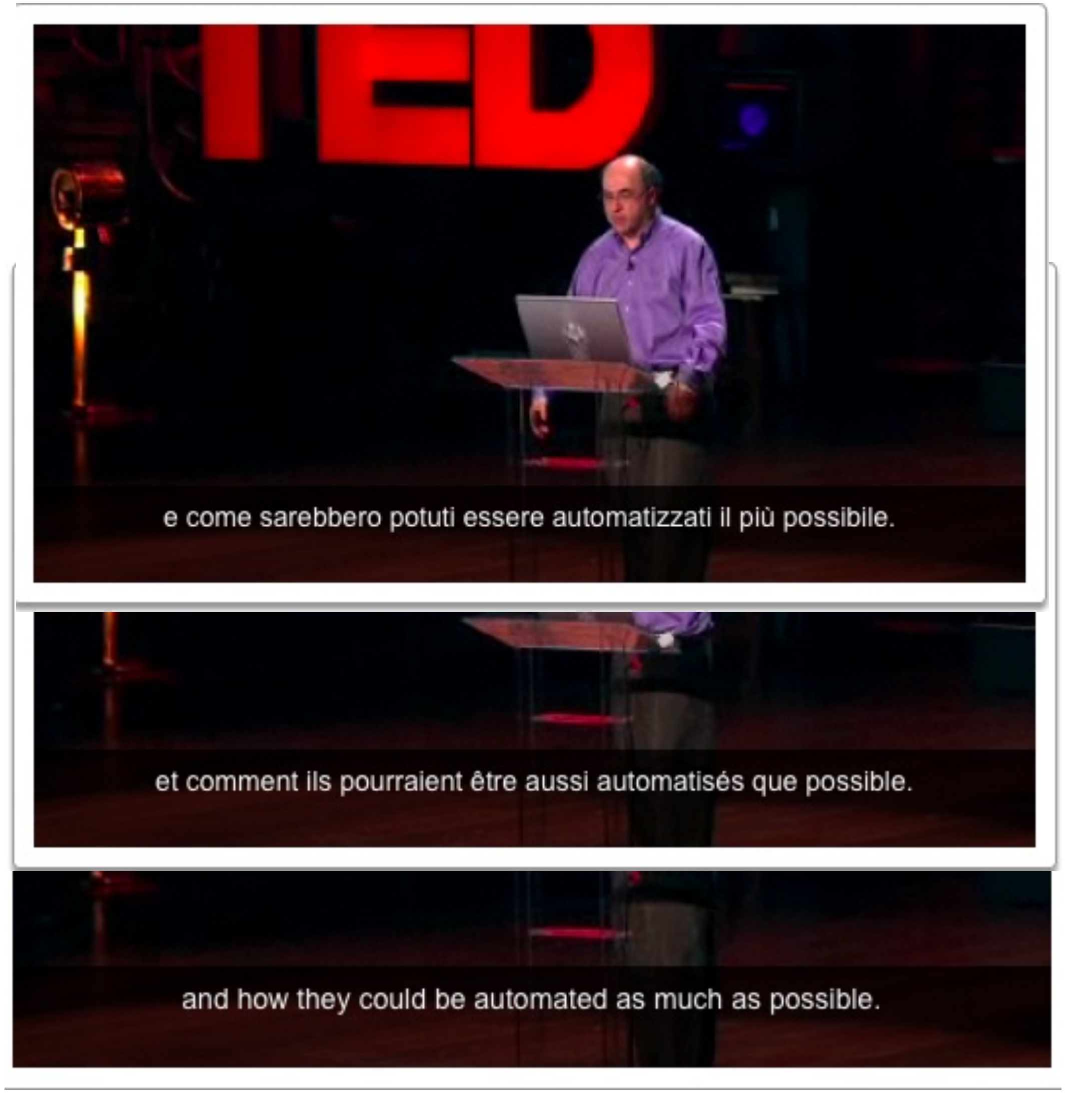
- TED LLC is non-profit
- Two annual events
- Short talks
- Variety of topics
- Website with:
  - Videos
  - Transcripts
  - Translations
- CC License

TED Talks Translations (from English)

	Nov '10	Nov '11	Nov '12	Nov '13	Nov '14
Talks (EN)	800	1,080	1,395	~1650	1875
Languages	80	83	93	103	105
Translators	4,000	6,823	8,382	11,010	18,699
Translations	12,500	24,287 +94%	32,707 +34%	49,607 +52%	65,290 +32%


Talks available at TED site (Nov 2014)
Turkish
Spanish
Slovenian
Russian
Romanian
Portuguese (Brazilian)
Polish
Italian
Hebrew
German
French (France)
Farsi/Persian
English
Dutch
Chinese (traditional)
Arabic
0
500
1000
1500
2000

Human task: subtitling and translating



e come sarebbero potuti essere automatizzati il più possibile.

et comment ils pourraient être aussi automatisés que possible.

and how they could be automated as much as possible.

✓ 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 and 2014

- **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, ...**
  - **Few in-domain training data for GER, IT: untranscribed**
- **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**

2014 Tracks

- **Automatic Speech Recognition (ASR)**
  - Transcription of talks from audio to text
  - English (TED), German (TEDX), **Italian (TEDX)**
- **Spoken Language Translation (SLT)**
  - Translation of talks from audio (or ASR output) to text
  - English-French, German<->English, **Italian<->English**
  - English-Arabic, English-Chinese **unofficial pairs**
- **Machine Translation (MT)**
  - Translation of talks from text to text
  - English-French, German<->English, **Italian<->English**
  - + X-English and English-X **12 unofficial pairs**

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

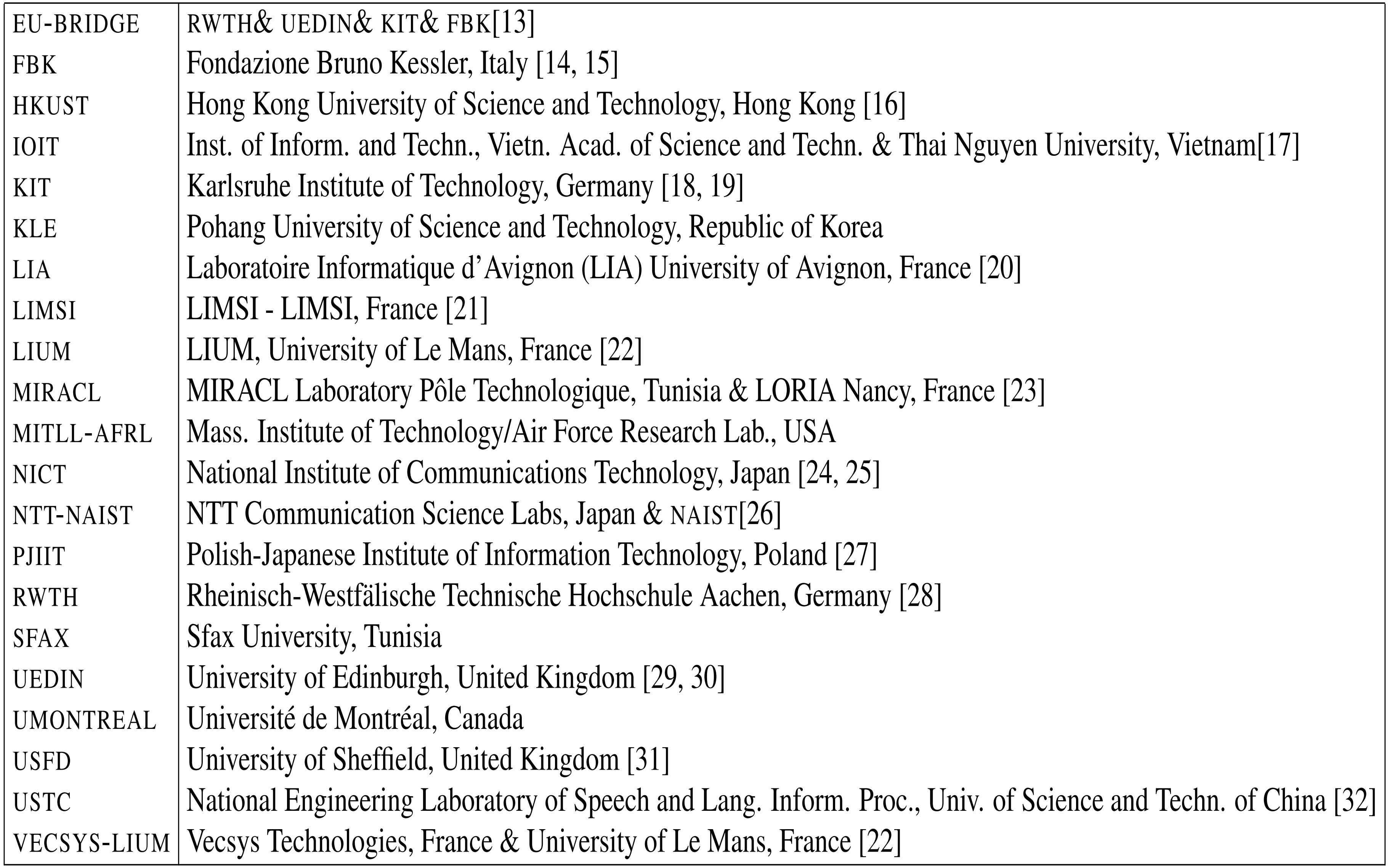
Specifications

Conditions	ASR	SLT	MT
Input: Pre-segmented	no	yes	yes
Input: Cased & Punctuated		no	yes
Output: Cased & Punctuated	no	yes	yes
Automatic evaluation [1]	yes	yes	yes
Human eval (En-Fr/De)			yes

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

[1] Prepared non trivial reference baselines for all MT directions.

Participants



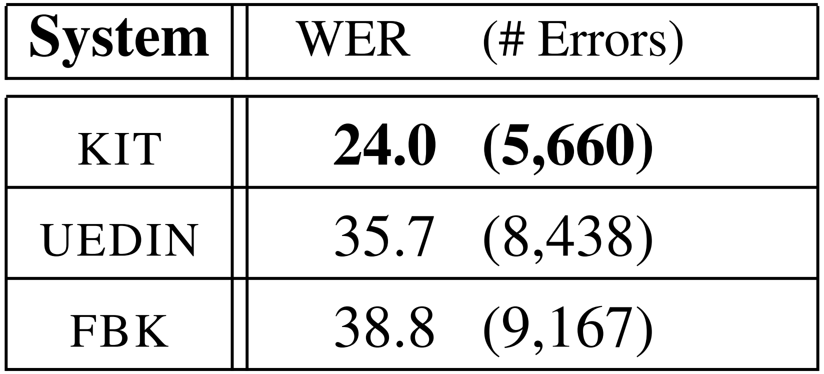
EU-BRIDGE	RWTH& UEDIN& KIT& FBK[13]
FBK	Fondazione Bruno Kessler, Italy [14, 15]
HKUST	Hong Kong University of Science and Technology, Hong Kong [16]
IOIT	Inst. of Inform. and Techn., Vietn. Acad. of Science and Techn. & Thai Nguyen University, Vietnam[17]
KIT	Karlsruhe Institute of Technology, Germany [18, 19]
KLE	Pohang University of Science and Technology, Republic of Korea
LIA	Laboratoire Informatique d’Avignon (LIA) University of Avignon, France [20]
LIMSI	LIMSI - LIMSI, France [21]
LIUM	LIUM, University of Le Mans, France [22]
MIRACL	MIRACL Laboratory Pôle Technologique, Tunisia & LORIA Nancy, France [23]
MITLL-AFRL	Mass. Institute of Technology/Air Force Research Lab., USA
NICT	National Institute of Communications Technology, Japan [24, 25]
NTT-NAIST	NTT Communication Science Labs, Japan & NAIST[26]
PJIIT	Polish-Japanese Institute of Information Technology, Poland [27]
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen, Germany [28]
SFAX	Sfax University, Tunisia
UEDIN	University of Edinburgh, United Kingdom [29, 30]
UMONTREAL	Université de Montréal, Canada
USFD	University of Sheffield, United Kingdom [31]
USTC	National Engineering Laboratory of Speech and Lang. Inform. Proc., Univ. of Science and Techn. of China [32]
VECSYS-LIUM	Vecsys Technologies, France & University of Le Mans, France [22]

Results: ASR English (WER%)

Run	TST14	TST13	TST13 IWSLT13
NICT	8.4	10.6	13.5
EU-BRIDGE	9.8	-	-
MITLL-AFR	9.9	13.7	15.9
KIT	11.4	14.2	14.4
FBK	11.4	14.7	23.2
LIUM	12.3	16.0	-
UEDIN	12.7	16.3	22.1
IOIT	19.7	24.0	27.2

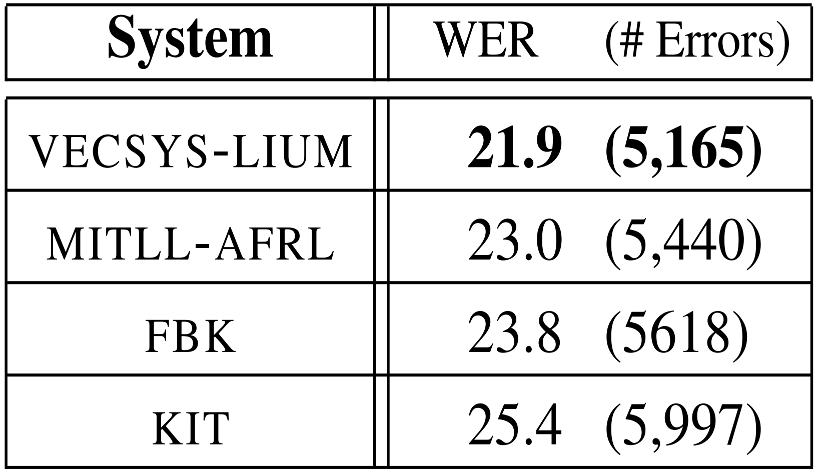
Results: ASR German and Italian

TEDX ASR German (ASR$_{DE}$)



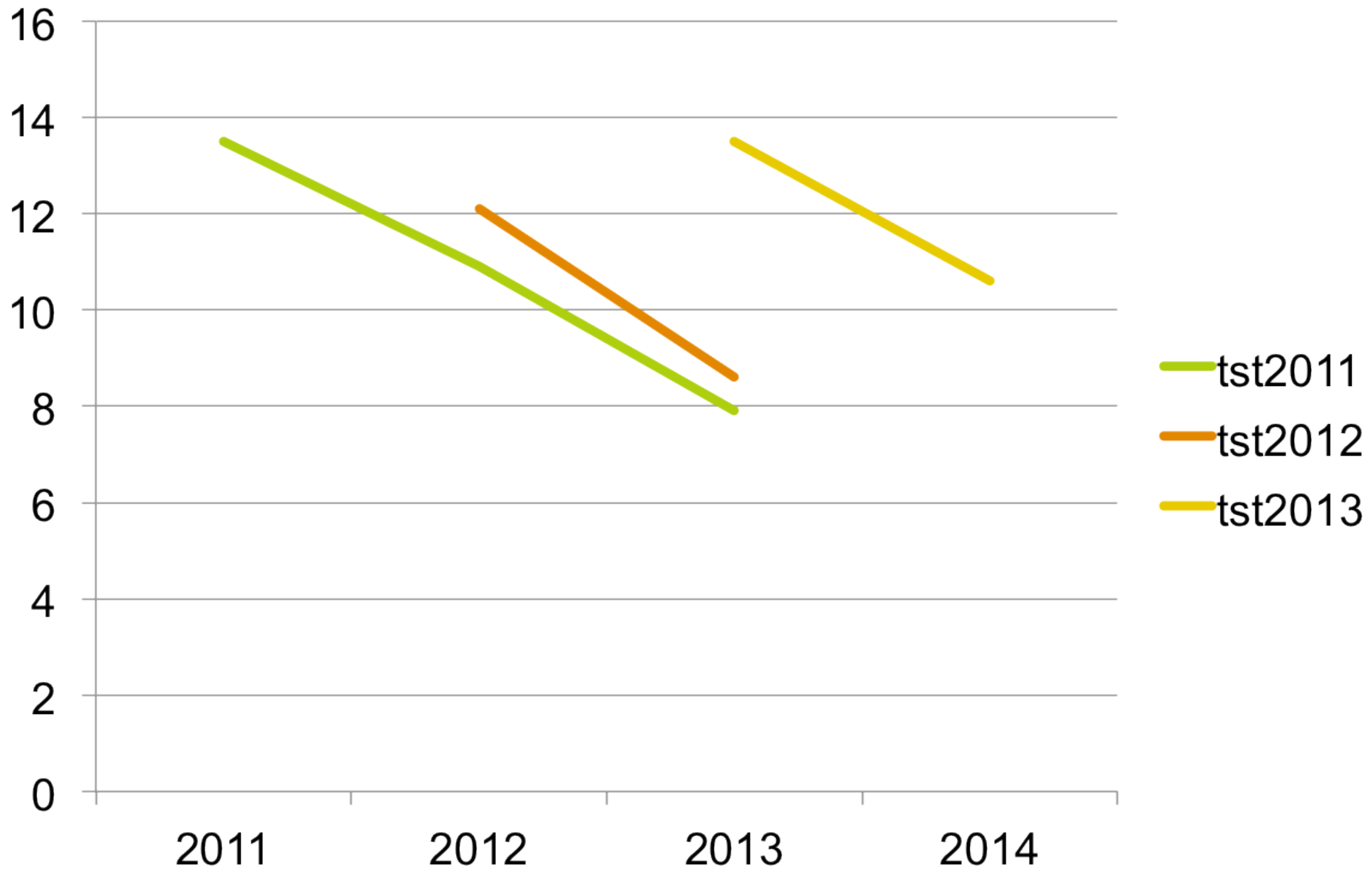
System	WER	(# Errors)
KIT	**24.0**	**(5,660)**
UEDIN	35.7	(8,438)
FBK	38.8	(9,167)

TEDX ASR Italian (ASR$_{IT}$)

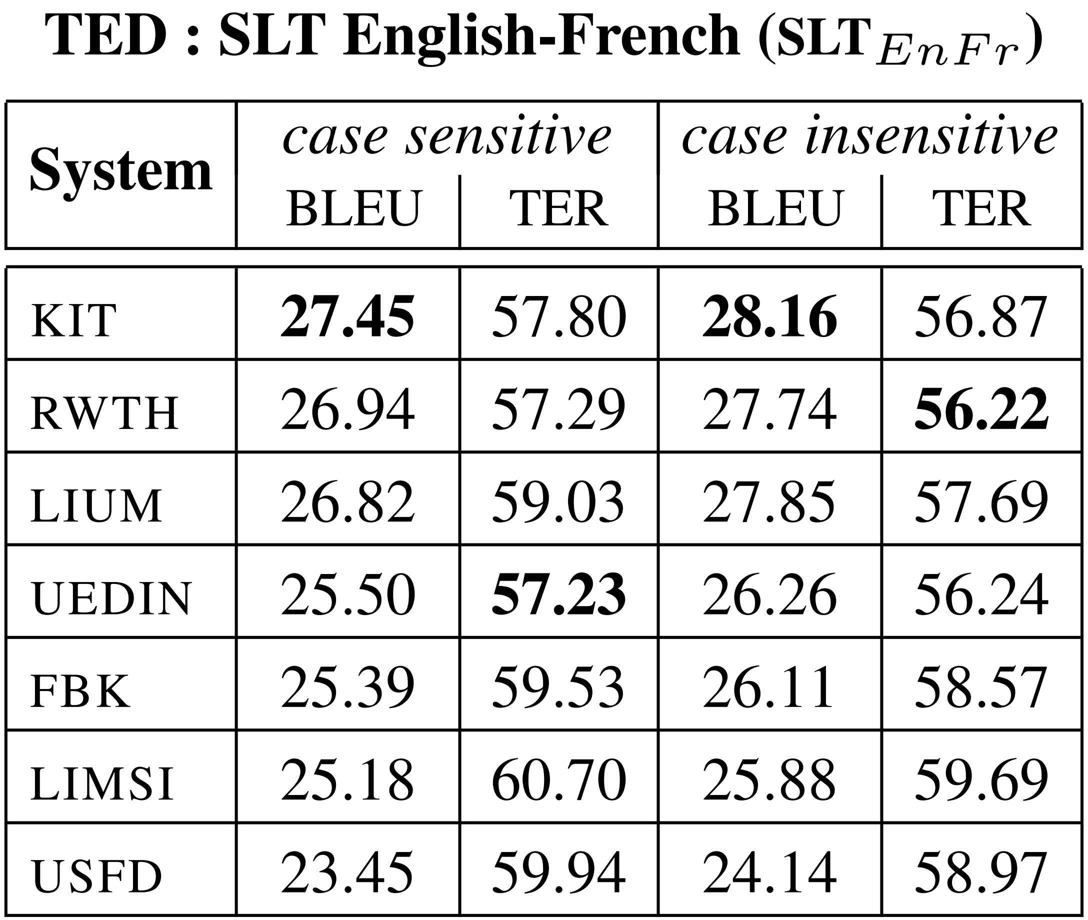


System	WER	(# Errors)
VECSYS-LIUM	**21.9**	**(5,165)**
MITLL-AFRL	23.0	(5,440)
FBK	23.8	(5618)
KIT	25.4	(5,997)

Progress in ASR En (best systems WER%)



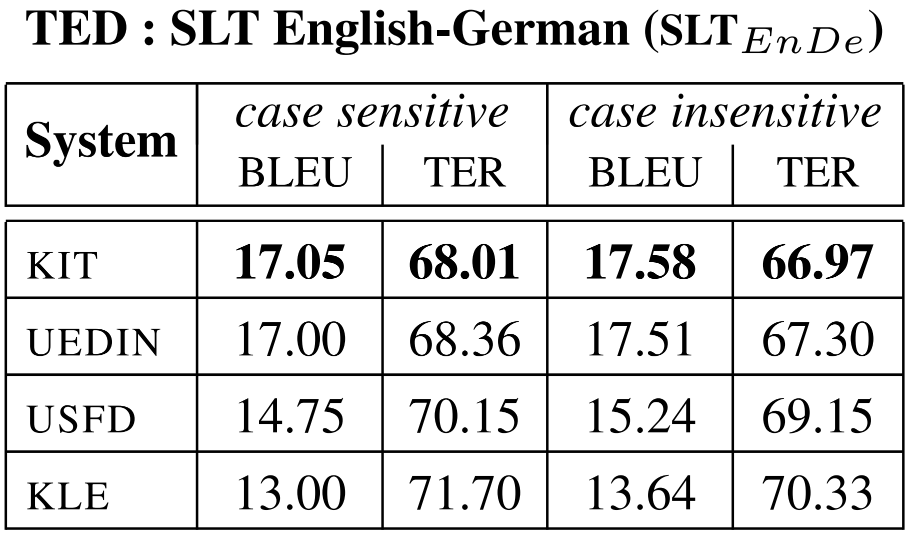
Results: SLT



TED : SLT English-French (SLT$_{EnFr}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
KIT	**27.45**	57.80	**28.16**	56.87
RWTH	26.94	57.29	27.74	**56.22**
LIUM	26.82	59.03	27.85	57.69
UEDIN	25.50	**57.23**	26.26	56.24
FBK	25.39	59.53	26.11	58.57
LIMSI	25.18	60.70	25.88	59.69
USFD	23.45	59.94	24.14	58.97

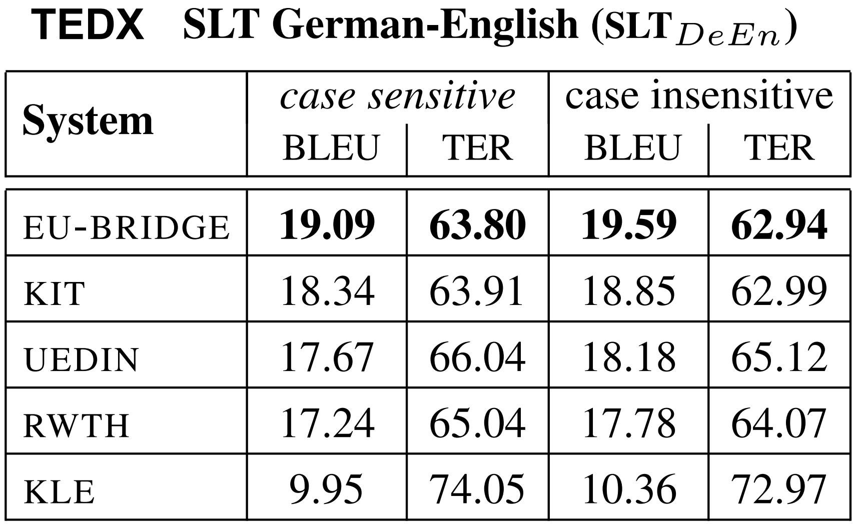
Results: SLT



TED : SLT English-German (SLT$_{EnDe}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**17.05**	**68.01**	**17.58**	**66.97**
UEDIN	17.00	68.36	17.51	67.30
USFD	14.75	70.15	15.24	69.15
KLE	13.00	71.70	13.64	70.33

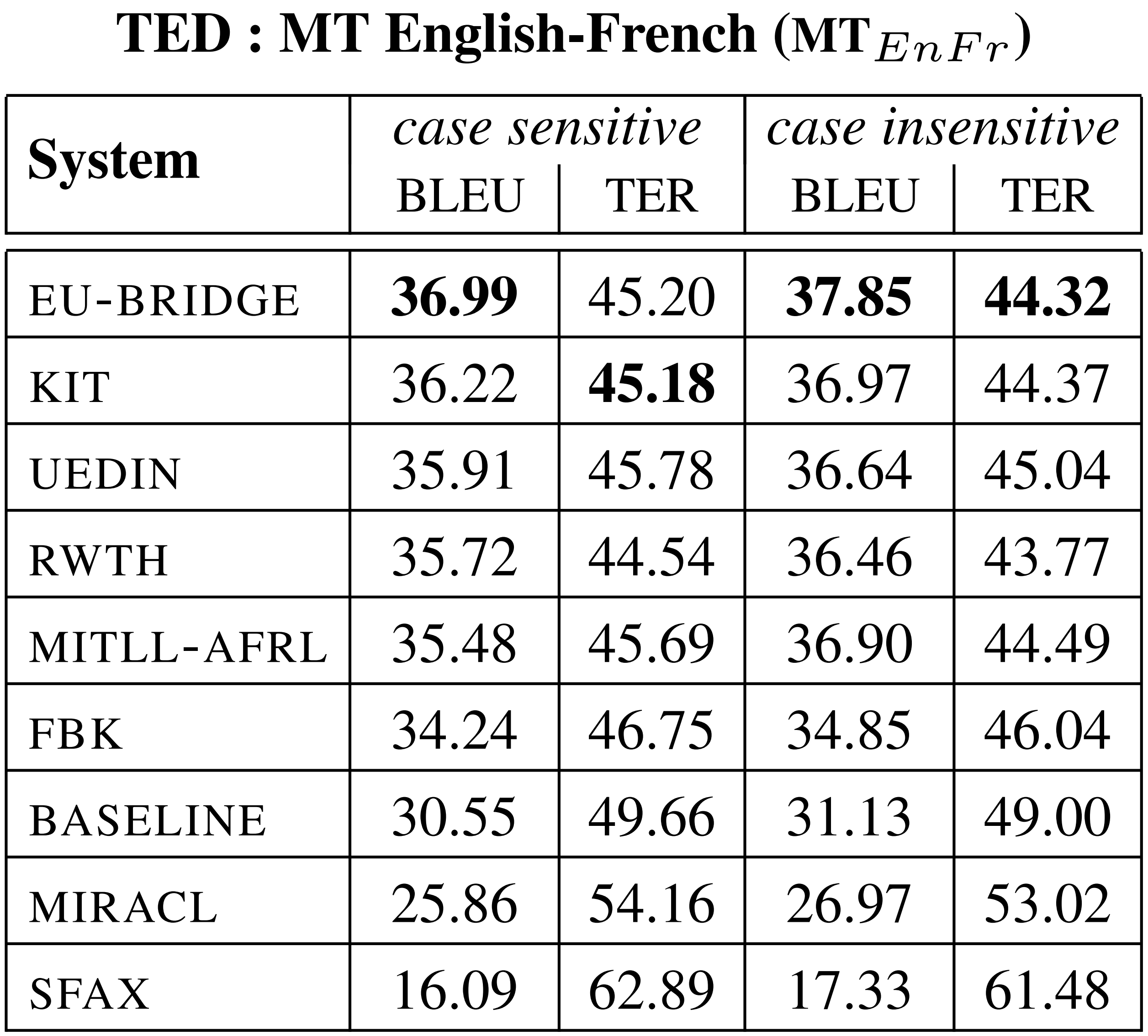
TEDX



SLT German-English (SLT$_{DeEn}$)

System	*case sensitive*		case insensitive	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**19.09**	**63.80**	**19.59**	**62.94**
KIT	18.34	63.91	18.85	62.99
UEDIN	17.67	66.04	18.18	65.12
RWTH	17.24	65.04	17.78	64.07
KLE	9.95	74.05	10.36	72.97

Results: MT

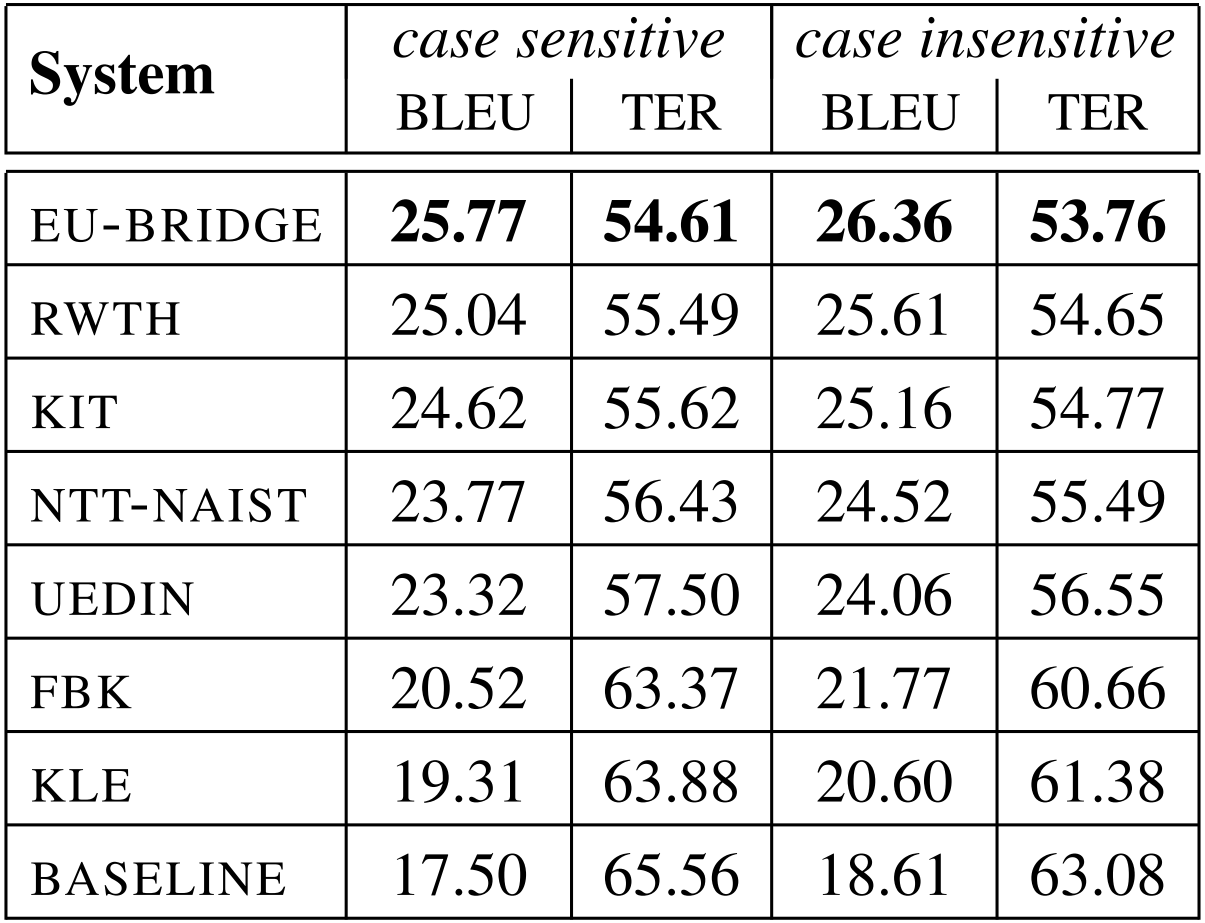


TED : MT English-French (MT$_{EnFr}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**36.99**	45.20	**37.85**	**44.32**
KIT	36.22	**45.18**	36.97	44.37
UEDIN	35.91	45.78	36.64	45.04
RWTH	35.72	44.54	36.46	43.77
MITLL-AFRL	35.48	45.69	36.90	44.49
FBK	34.24	46.75	34.85	46.04
BASELINE	30.55	49.66	31.13	49.00
MIRACL	25.86	54.16	26.97	53.02
SFAX	16.09	62.89	17.33	61.48

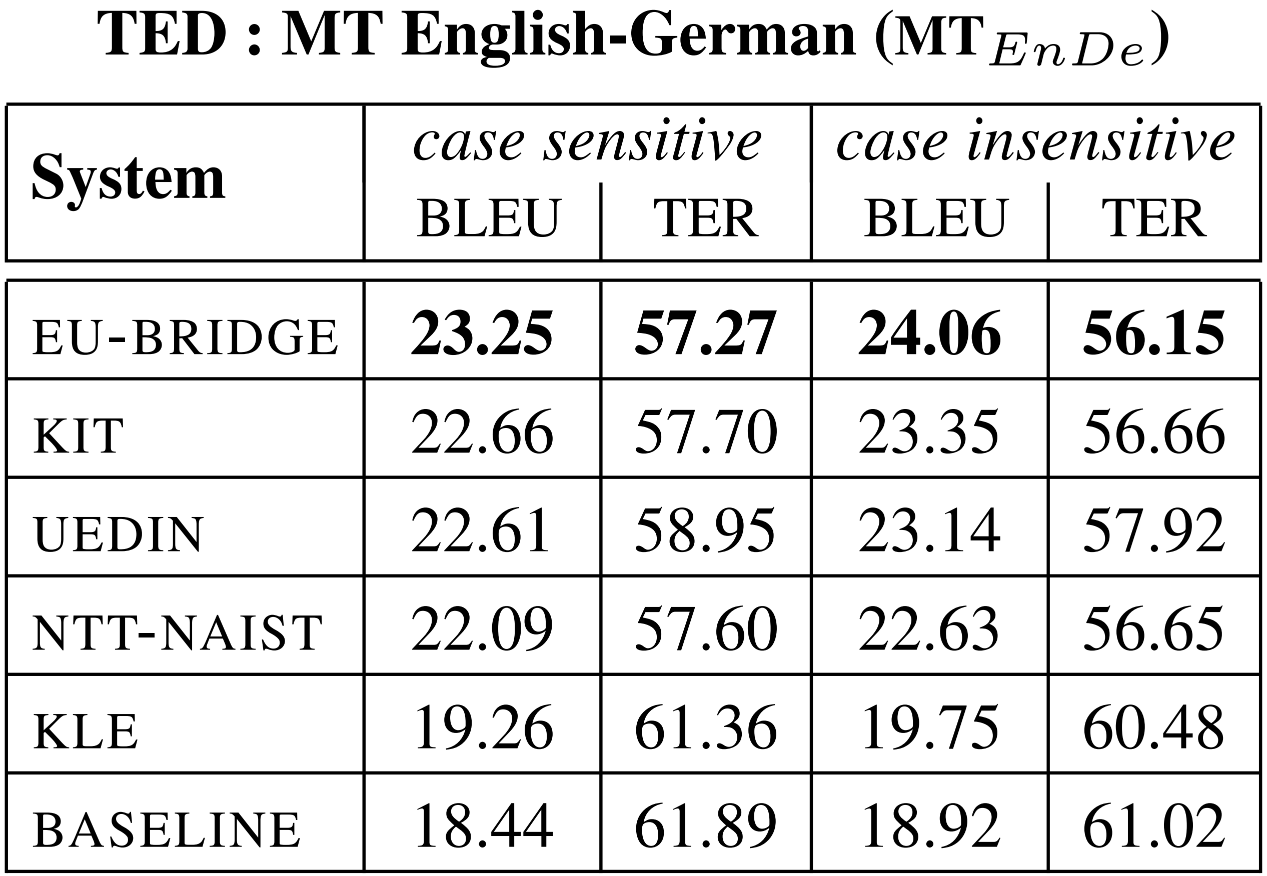
Results: MT

TEDX **MT German-English (SLT$_{DeEn}$)**



System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**25.77**	**54.61**	**26.36**	**53.76**
RWTH	25.04	55.49	25.61	54.65
KIT	24.62	55.62	25.16	54.77
NTT-NAIST	23.77	56.43	24.52	55.49
UEDIN	23.32	57.50	24.06	56.55
FBK	20.52	63.37	21.77	60.66
KLE	19.31	63.88	20.60	61.38
BASELINE	17.50	65.56	18.61	63.08

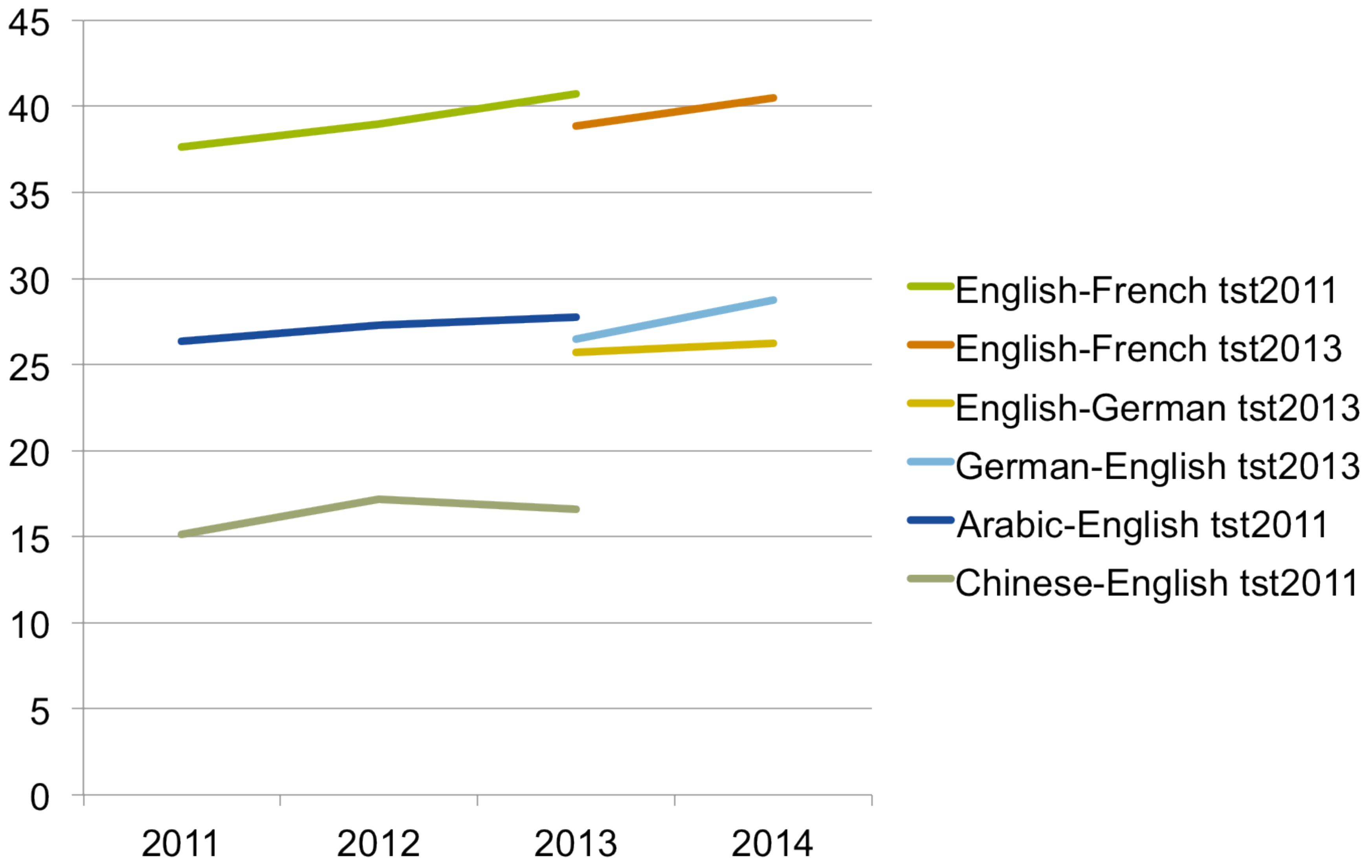
Results: MT



TED : MT English-German (MT$_{EnDe}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**23.25**	**57.27**	**24.06**	**56.15**
KIT	22.66	57.70	23.35	56.66
UEDIN	22.61	58.95	23.14	57.92
NTT-NAIST	22.09	57.60	22.63	56.65
KLE	19.26	61.36	19.75	60.48
BASELINE	18.44	61.89	18.92	61.02

Progress in MT (best systems BLEU%)



Human Evaluation

- Following IWSLT 2013: ***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
- Evaluation of **MT-*EnDe*** and **MT-*EnFr*** tracks
- Performed on 2013 progress test set (*tst2013*)

Evaluation Dataset

Human Evaluation (HE) Set:

- a subset of *tst2013*
  - initial 60% of the 16 different talks composing *tst2013*
  - ~11,000 words
- *EnDe*: 628 segments
- *EnFr*: 622 segments

Evaluation Setup

Lesson learned from IWSLT 2013:

- most informative and reliable HTER:
    - not by using the targeted reference only
    - but by exploiting all post-edits

Evaluation Setup

Lesson learned from IWSLT 2013:

- most informative and reliable HTER:
  - not by using the targeted reference only
  - but by exploiting all post-edits

SRC:
But why would you reconcile after a fight?

Targeted Reference Only

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: Mais pourquoi **se** **réconcilier** après un combat ?

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

TER:
23.33

Evaluation Setup

Lesson learned from IWSLT 2013:

- most informative and reliable HTER:
  - not by using the targeted reference only
  - but by exploiting all post-edits

IWSLT 2014 official evaluation:

- HTER calculated on multiple references (post-edits)
  - *EnDe:* 5 participants => 5 post-edits
  - EnFr: 7 participants => 5 post-edits

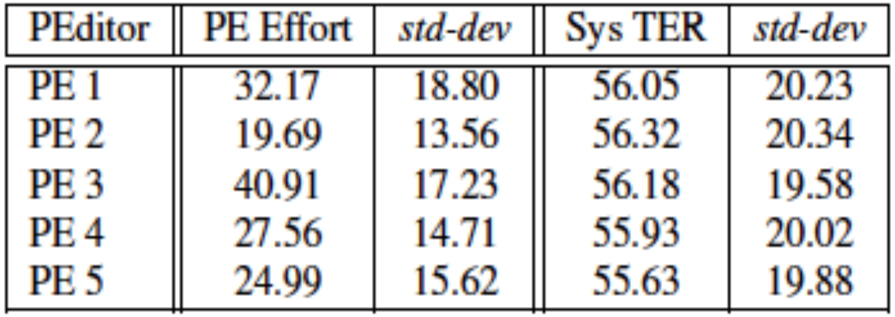
Data Collection

- *Bilingual* Post-Editing
  - professional translators were required to post-edit the MT output directly according to the source sentence
- Data preparation:
  - 5 systems p-edited by 5 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

Collected Data

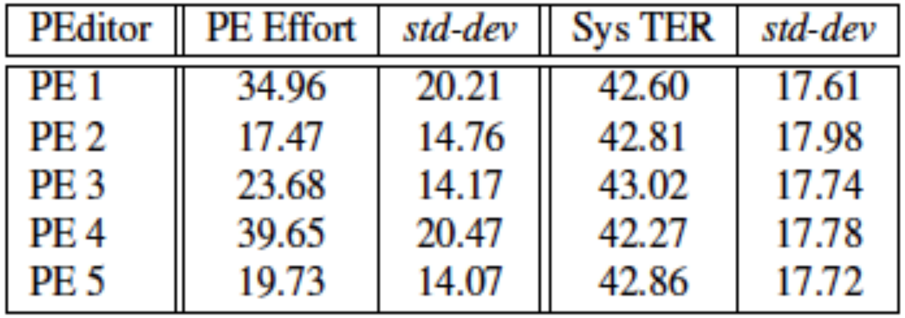
- Collected Post-edits
  - 5 new references for each sentence in the HE set
- Post-editors characteristics:

EnDe



PEditor	PE Effort	*std-dev*	Sys TER	*std-dev*
PE 1	32.17	18.80	56.05	20.23
PE 2	19.69	13.56	56.32	20.34
PE 3	40.91	17.23	56.18	19.58
PE 4	27.56	14.71	55.93	20.02
PE 5	24.99	15.62	55.63	19.88

EnFr

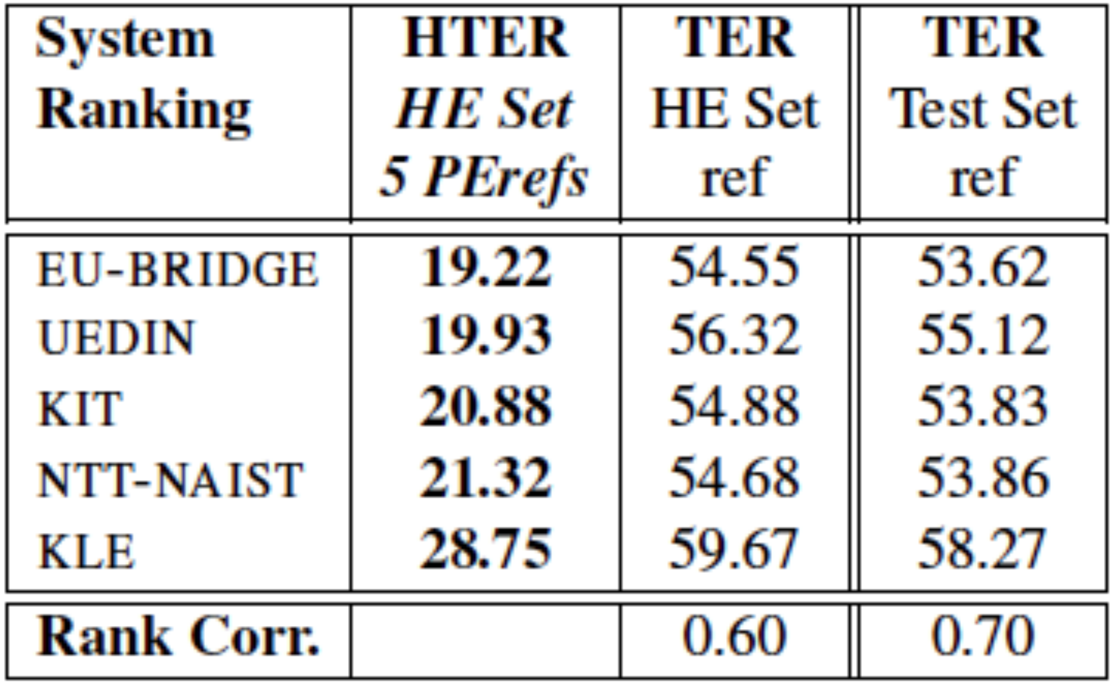


PEditor	PE Effort	*std-dev*	Sys TER	*std-dev*
PE 1	34.96	20.21	42.60	17.61
PE 2	17.47	14.76	42.81	17.98
PE 3	23.68	14.17	43.02	17.74
PE 4	39.65	20.47	42.27	17.78
PE 5	19.73	14.07	42.86	17.72

- PE effort (HTER): highly variable among post-editors
- MT outputs assigned to translators (Sys TER): very homogeneous

Evaluation Results - *EnDe*

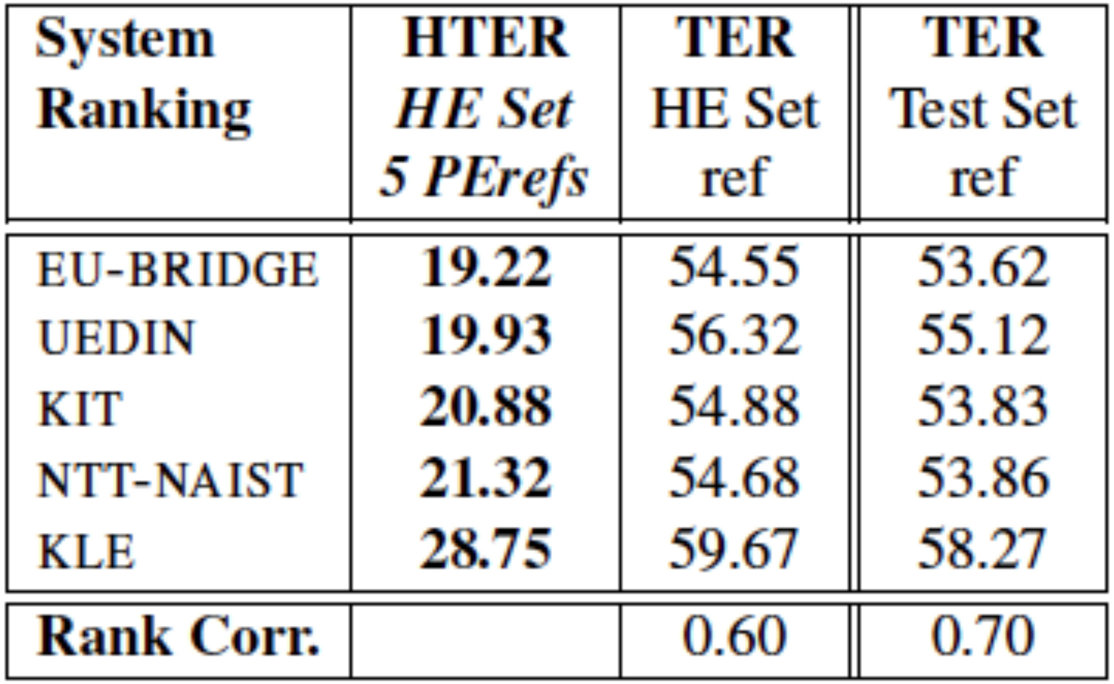
- HTER calculated on all 5 post-edits available
  - including targeted translation



System Ranking	**HTER** ***HE Set*** ***5 PErefs***	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**19.22**	54.55	53.62
UEDIN	**19.93**	56.32	55.12
KIT	**20.88**	54.88	53.83
NTT-NAIST	**21.32**	54.68	53.86
KLE	**28.75**	59.67	58.27
Rank Corr.		0.60	0.70

Evaluation Results - *EnDe*

- HTER calculated on all 5 post-edits available
  - including targeted translation

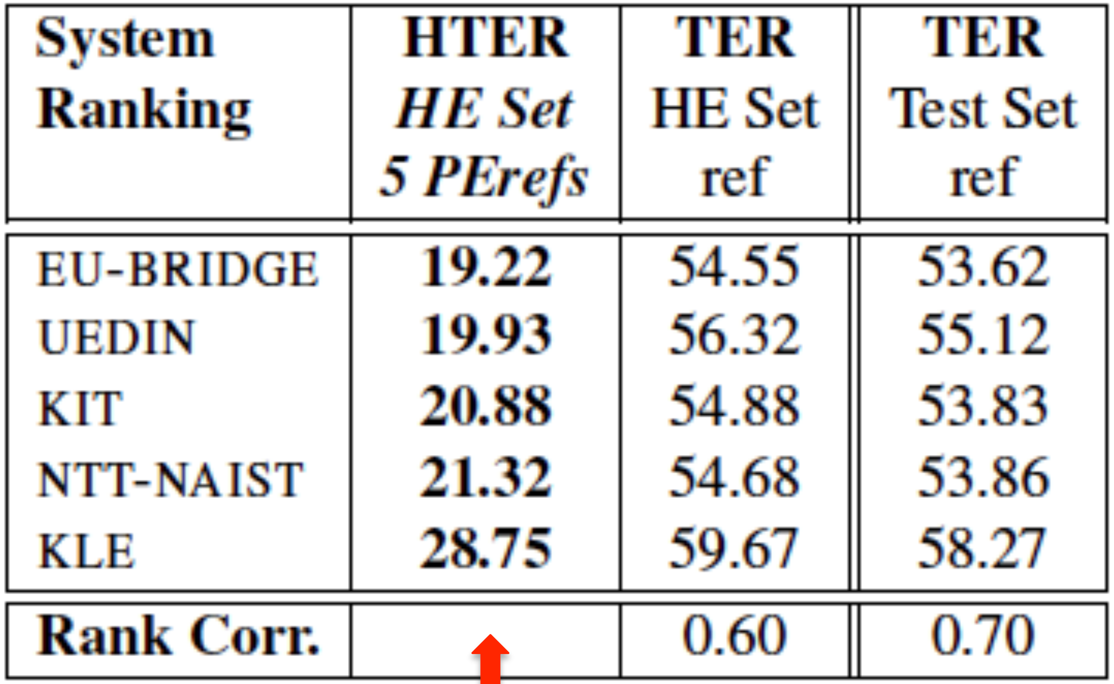


System Ranking	**HTER** ***HE Set*** ***5 PErefs***	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**19.22**	54.55	53.62
UEDIN	**19.93**	56.32	55.12
KIT	**20.88**	54.88	53.83
NTT-NAIST	**21.32**	54.68	53.86
KLE	**28.75**	59.67	58.27
Rank Corr.		0.60	0.70

-60%

Evaluation Results - *EnDe*

- HTER calculated on all 5 post-edits available
  - including targeted translation



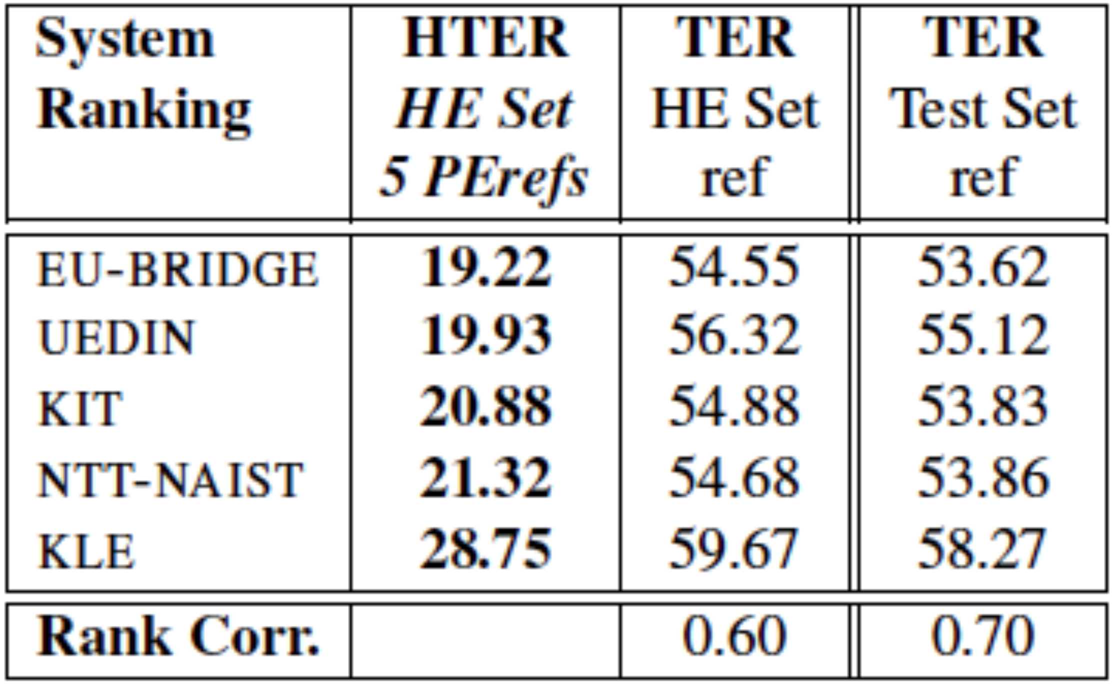
System Ranking	**HTER** ***HE Set*** ***5 PErefs***	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**19.22**	54.55	53.62
UEDIN	**19.93**	56.32	55.12
KIT	**20.88**	54.88	53.83
NTT-NAIST	**21.32**	54.68	53.86
KLE	**28.75**	59.67	58.27
Rank Corr.		0.60	0.70

Statistical Significance (Approximate Randomization):

Only KLE is significantly worse than all other systems at $p < 0.01$

Evaluation Results - *EnDe*

- HTER calculated on all 5 post-edits available
  - including targeted translation

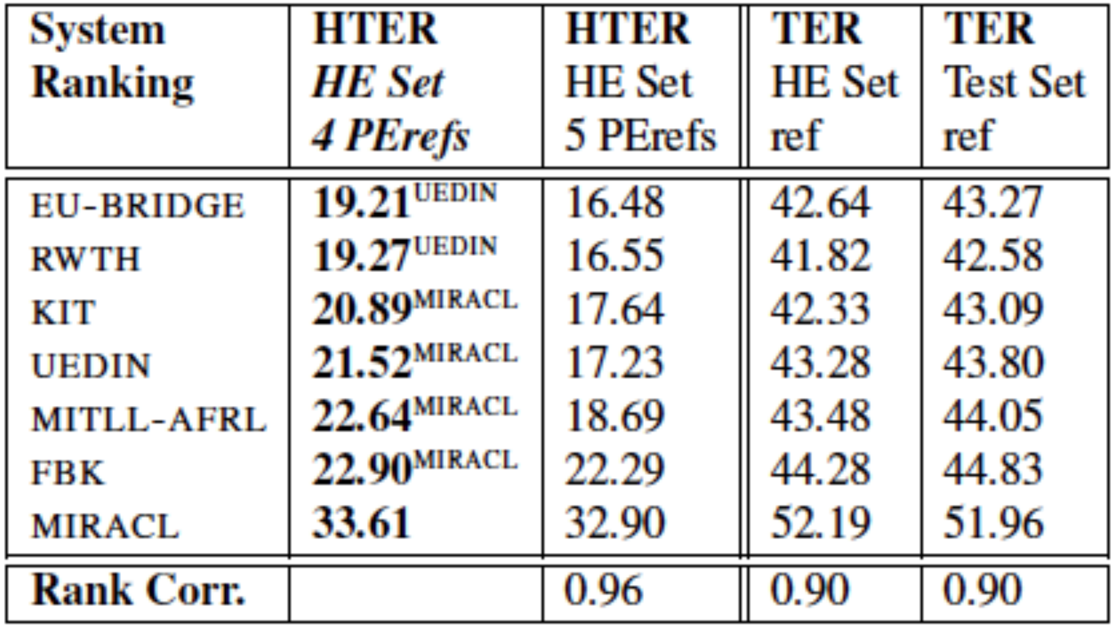


System	**HTER**	**TER**	**TER**
Ranking	***HE Set***	HE Set	Test Set
	5 PErefs	ref	ref
EU-BRIDGE	**19.22**	54.55	53.62
UEDIN	**19.93**	56.32	55.12
KIT	**20.88**	54.88	53.83
NTT-NAIST	**21.32**	54.68	53.86
KLE	**28.75**	59.67	58.27
Rank Corr.		0.60	0.70

Spearman's Rank Coefficient

Evaluation Results - *EnFr*

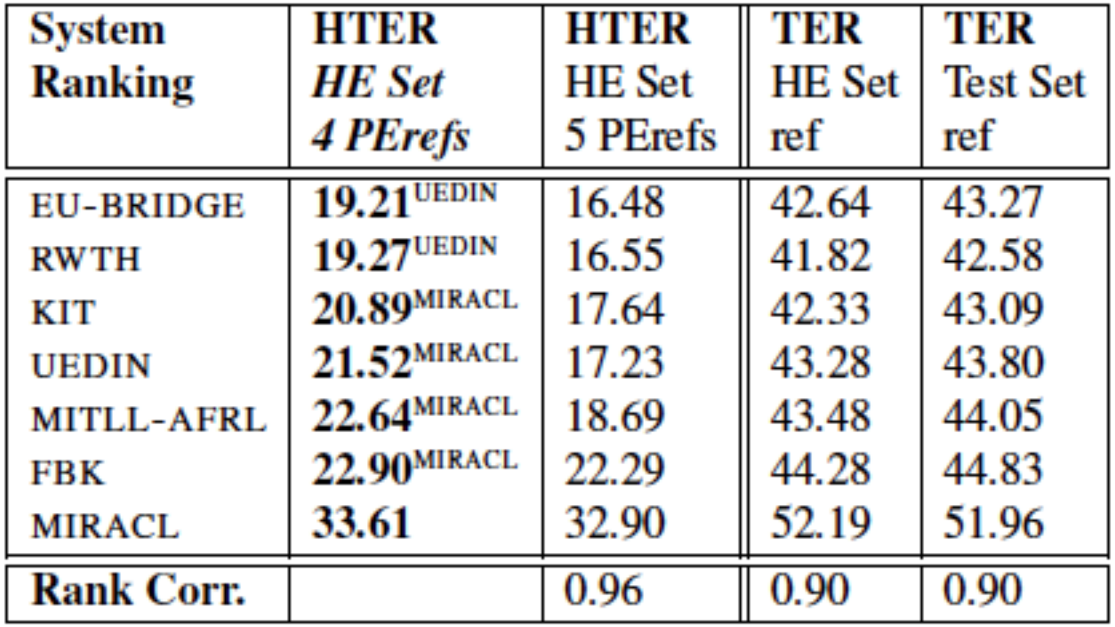
- HTER calculated on 4 post-edits:
  - systems 1-5: excluding system's targeted translation
  - systems 6-7: combination of the four post-edits which gave the best results



System	**HTER**	**HTER**	**TER**	**TER**
Ranking	***HE Set***	HE Set	HE Set	Test Set
	4 PErefs	5 PErefs	ref	ref
EU-BRIDGE	**19.21**$^{\text{UEDIN}}$	16.48	42.64	43.27
RWTH	**19.27**$^{\text{UEDIN}}$	16.55	41.82	42.58
KIT	**20.89**$^{\text{MIRACL}}$	17.64	42.33	43.09
UEDIN	**21.52**$^{\text{MIRACL}}$	17.23	43.28	43.80
MITLL-AFRL	**22.64**$^{\text{MIRACL}}$	18.69	43.48	44.05
FBK	**22.90**$^{\text{MIRACL}}$	22.29	44.28	44.83
MIRACL	**33.61**	32.90	52.19	51.96
Rank Corr.		0.96	0.90	0.90

Evaluation Results - *EnFr*

- HTER calculated on 4 post-edits:
    - systems 1-5: excluding system's targeted translation
    - systems 6-7: combination of the four post-edits which gave the best results

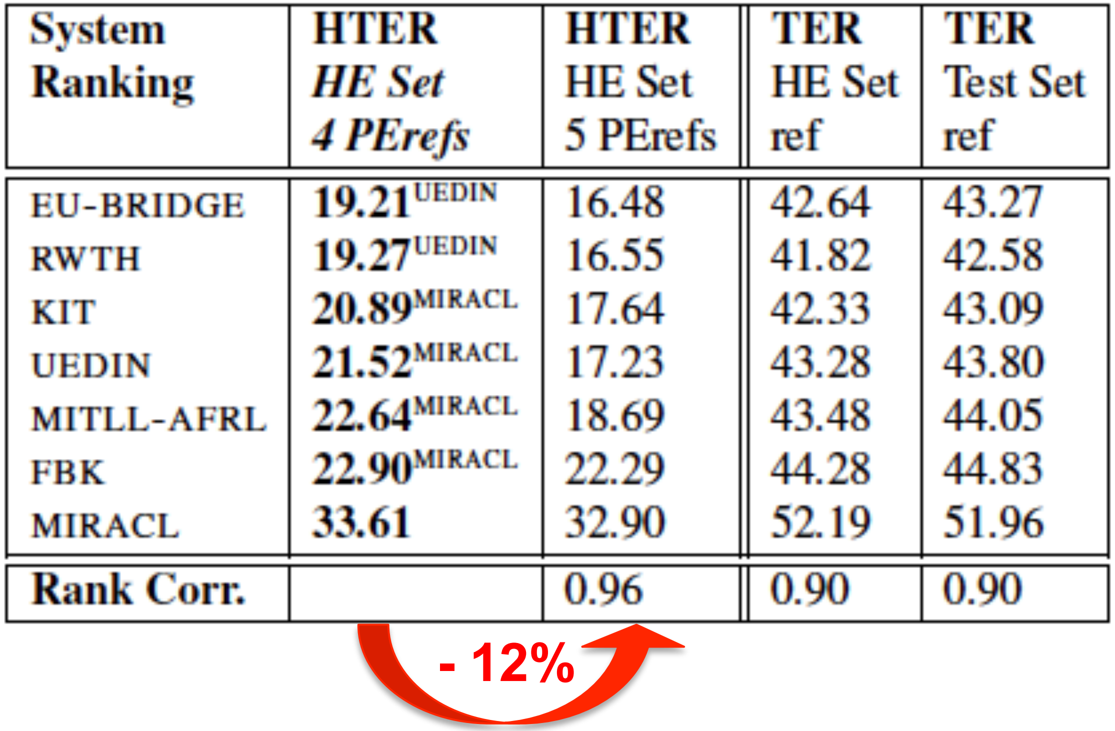


System	**HTER**	**HTER**	**TER**	**TER**
Ranking	***HE Set***	HE Set	HE Set	Test Set
	4 PErefs	5 PErefs	ref	ref
EU-BRIDGE	**19.21**$^{\text{UEDIN}}$	16.48	42.64	43.27
RWTH	**19.27**$^{\text{UEDIN}}$	16.55	41.82	42.58
KIT	**20.89**$^{\text{MIRACL}}$	17.64	42.33	43.09
UEDIN	**21.52**$^{\text{MIRACL}}$	17.23	43.28	43.80
MITLL-AFRL	**22.64**$^{\text{MIRACL}}$	18.69	43.48	44.05
FBK	**22.90**$^{\text{MIRACL}}$	22.29	44.28	44.83
MIRACL	**33.61**	32.90	52.19	51.96
Rank Corr.		0.96	0.90	0.90

-50%

Evaluation Results - *EnFr*

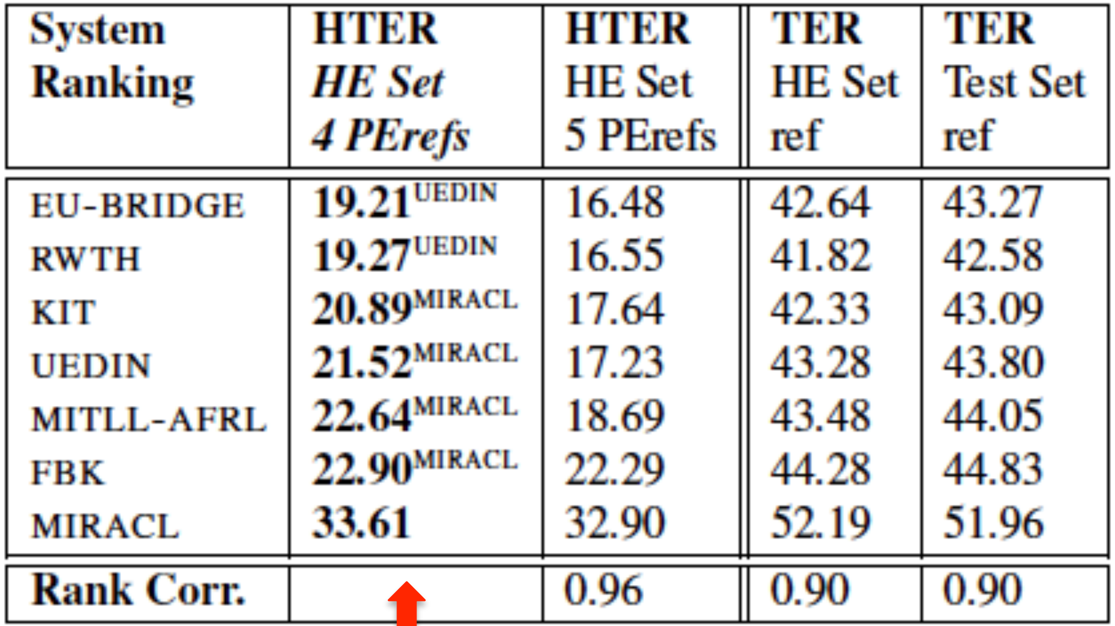
- HTER calculated on 4 post-edits:
  - systems 1-5: excluding system's targeted translation
  - systems 6-7: combination of the four post-edits which gave the best results



System	**HTER**	**HTER**	**TER**	**TER**
Ranking	***HE Set***	HE Set	HE Set	Test Set
	4 PErefs	5 PErefs	ref	ref
EU-BRIDGE	**19.21**^{UEDIN}	16.48	42.64	43.27
RWTH	**19.27**^{UEDIN}	16.55	41.82	42.58
KIT	**20.89**^{MIRACL}	17.64	42.33	43.09
UEDIN	**21.52**^{MIRACL}	17.23	43.28	43.80
MITLL-AFRL	**22.64**^{MIRACL}	18.69	43.48	44.05
FBK	**22.90**^{MIRACL}	22.29	44.28	44.83
MIRACL	**33.61**	32.90	52.19	51.96
Rank Corr.		0.96	0.90	0.90

Evaluation Results - *EnFr*

- HTER calculated on 4 post-edits:
  - systems 1-5: excluding system's targeted translation
  - systems 6-7: combination of the four post-edits which gave the best results

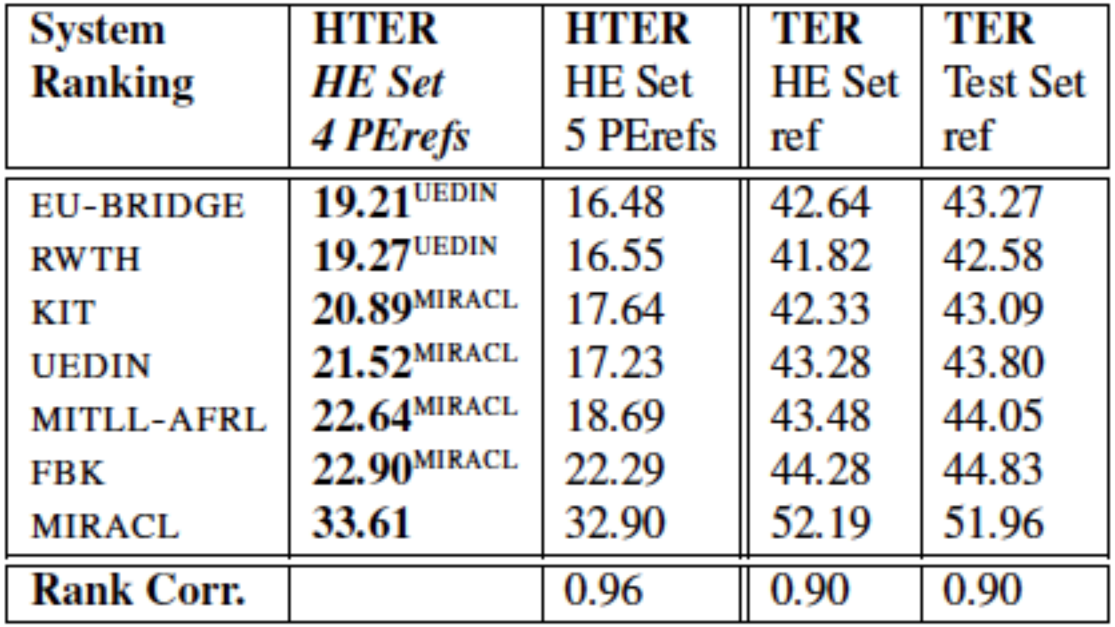


System	**HTER**	**HTER**	**TER**	**TER**
Ranking	***HE Set***	HE Set	HE Set	Test Set
	4 PErefs	5 PErefs	ref	ref
EU-BRIDGE	**19.21^{UEDIN}**	16.48	42.64	43.27
RWTH	**19.27^{UEDIN}**	16.55	41.82	42.58
KIT	**20.89^{MIRACL}**	17.64	42.33	43.09
UEDIN	**21.52^{MIRACL}**	17.23	43.28	43.80
MITLL-AFRL	**22.64^{MIRACL}**	18.69	43.48	44.05
FBK	**22.90^{MIRACL}**	22.29	44.28	44.83
MIRACL	**33.61**	32.90	52.19	51.96
Rank Corr.		0.96	0.90	0.90

Statistical Significance (Approximate Randomization) at $p < 0.01$:

Evaluation Results - *EnFr*

- HTER calculated on 4 post-edits:
  - systems 1-5: excluding system’s targeted translation
  - systems 6-7: combination of the four post-edits which gave the best results



System Ranking	**HTER** ***HE Set 4 PErefs***	**HTER** HE Set 5 PErefs	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**19.21**$^{\text{UEDIN}}$	16.48	42.64	43.27
RWTH	**19.27**$^{\text{UEDIN}}$	16.55	41.82	42.58
KIT	**20.89**$^{\text{MIRACL}}$	17.64	42.33	43.09
UEDIN	**21.52**$^{\text{MIRACL}}$	17.23	43.28	43.80
MITLL-AFRL	**22.64**$^{\text{MIRACL}}$	18.69	43.48	44.05
FBK	**22.90**$^{\text{MIRACL}}$	22.29	44.28	44.83
MIRACL	**33.61**	32.90	52.19	51.96
Rank Corr.		0.96	0.90	0.90

Spearman’s Rank Coefficient

Future plans

- Add more ASR languages
- Extend the concept of language experts, more help in scoring and normalization
- Include more English to X translation tasks for MT and SLT
  - Target Asian languages such as Japanese, Korean, Thai, Vietnamese,
- Ask participants to provide ASR real-time factor
- Add additional track based on tourist domain
  - Coordinated by NICT
- Continue with HE based on post-editing
  - Funding by H2020 CSA Cracker

Credits

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

Questions?