Departament de Traducció
i d'Interpretació i d'Estudis
de l'Àsia Oriental
UAB
Universitat Autònoma de Barcelona

Machine translation and audio description: is it worth it? Assessing post-editing effort

Anna Fernández-Torné & Anna Matamala

Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental/ CAIAC, UAB/ TransMedia Catalonia

anna.torne@gmail.com
anna.matamala@uab.cat

Languages and the Media Conference 2014
Berlin, 5-7 November 2014

Funded by FFI2012-31024 (Spanish Ministerio de Economía y Competitividad) & 2014SGR27 (Catalan Government)

The ALST project

- 2013-2015
- Funded by the Spanish Ministerio de Economía y Competitividad
- Main researcher: Anna Matamala (UAB)
- PhD students: Anna Fernández-Torné & Carla Ortiz-Boix
- Team members from UAB (8) and external universities (6).

Aims

- Apply technologies
  - Speech recognition
  - Machine translation
  - Speech synthesis

…to make content accessible

a) audio description

b) voice-over

Why?

- Three technologies:
  - can they help translators/describers?
  - increase accessibility?
  - future integration in semi-automatic work flows?
  - in all environments? Professional/amateurs?

Does it make sense?

- Limited funding > rely on free online resources

- Existing projects & previous work: SUMAT, Eu-Bridge, research by Martin Volk, DCU researchers, Specia & De Sousa, etc.

- Critical voices from practitioners… although some voices from the industry in favour

It does (hopefully)!

- Exploratory
- Oral transfer modes (not subtitles): AD (sensorial accessibility) & VO (linguistic accessibility)> narrator with a clear voicing
- English, Spanish… and Catalan
- Focus on the describer/translator: (perceived) postediting effort

ALST Work distribution

- SR in VO/ documentaries (2015, with Pablo Romero & Lukasz Daniluk: experimental design)
- SR in AD (end 2014, with H. Delgado & J. Serrano)

- MT in VO (Carla Ortiz, tests done)
- MT in AD (Anna Fernández-Torné, tests done)

- TTS VO (2015)
- TTS AD (2013-2014: Dubrovnik 2014, Jostrans)

MT in AD

- Effort in
  - Creating an AD
  - Translating an English AD into Catalan (human translation)
  - Postediting a MT AD
- Final quality (future work)

MT in AD

- Pre-test to select the machine translation engine (focus of this presentation)

- Main experiment (finished, ongoing analysis)
  - 12 participants (selection criteria)
  - 3 tasks: creation, human translation, post-editing
  - Measures:
    - Subjective (perceived effort): questionnaires pre/post
    - Objective: time, keyboard logging data

MT in AD: the pretest

- Selection of the main free online engine (EN>CA) for AD
  - Apertium, Bing, Lucy, Google, Yandex
- 5 participants (professional translators)
- 5 MT versions of 14 AD units (“Closer”), 240 words, 3.09 min (randomized and balanced)

MT in AD: the pretest

- Post-editing effort
  - Automatic measures: postediting time & HTER (human-targeted translation edit rate).
  - 5-point Likert scale on PE difficulty and PE necessity
- MT output
  - Adequacy and fluency
  - Ranking of raw MT output

MT in AD: the pretest

- Why these measures?

  Previous research: ranking task (Avramidis et al 2012, Federmann 2012), expected and perceived post-editing effort annotations using adjectival interval-scales (Specia et al 2009, De Sousa et al 2011), HTER & postediting time (Specia 2011), error classification (Avramidis 2012, Popovic et al 2013), fluency and adequacy judgements (using interval-scales: Callison-Burch 2012; using Likert scales: Graham et al. 2013), etc.

MT in AD: the pretest

- Tool?
  - Analysis of existing tools: Casmacat, Addicter, SymEval, Appraise, Accept, TransCenter, PET
  - No tools include video and audio options
  - PET selected (no synchronisation was required at this stage) because (a) standalone tool, and (b) customisable (thanks to Wilker Aziz & Lucia Specia)

MT in AD: the pretest

- Instructions & brief training (PET)
- Watch dubbed version in Catalan with embedded English subtitles of the AD
- Instructions on postediting (as if this was a standard translation with no synchronisation or time-coding)

MT in AD: the pretest

- Post-editing instructions (inspired by O'Brien 2010, De Sousa et al 2011, Specia 2011, TAUS 2013, Housley 2012).
  - Perform the minimum amount of editing necessary to make the AD translation ready for voicing retaining as much raw translation as possible
  - Aim for grammatically, syntactically and semantically correct translation
  - Ensure that no information has been accidentally added or omitted
  - Ensure that the message transferred is acurate
  - Ensure that key terminology is correctly translated
  - Basic rules regarding spelling, punctuation and hyphenation apply

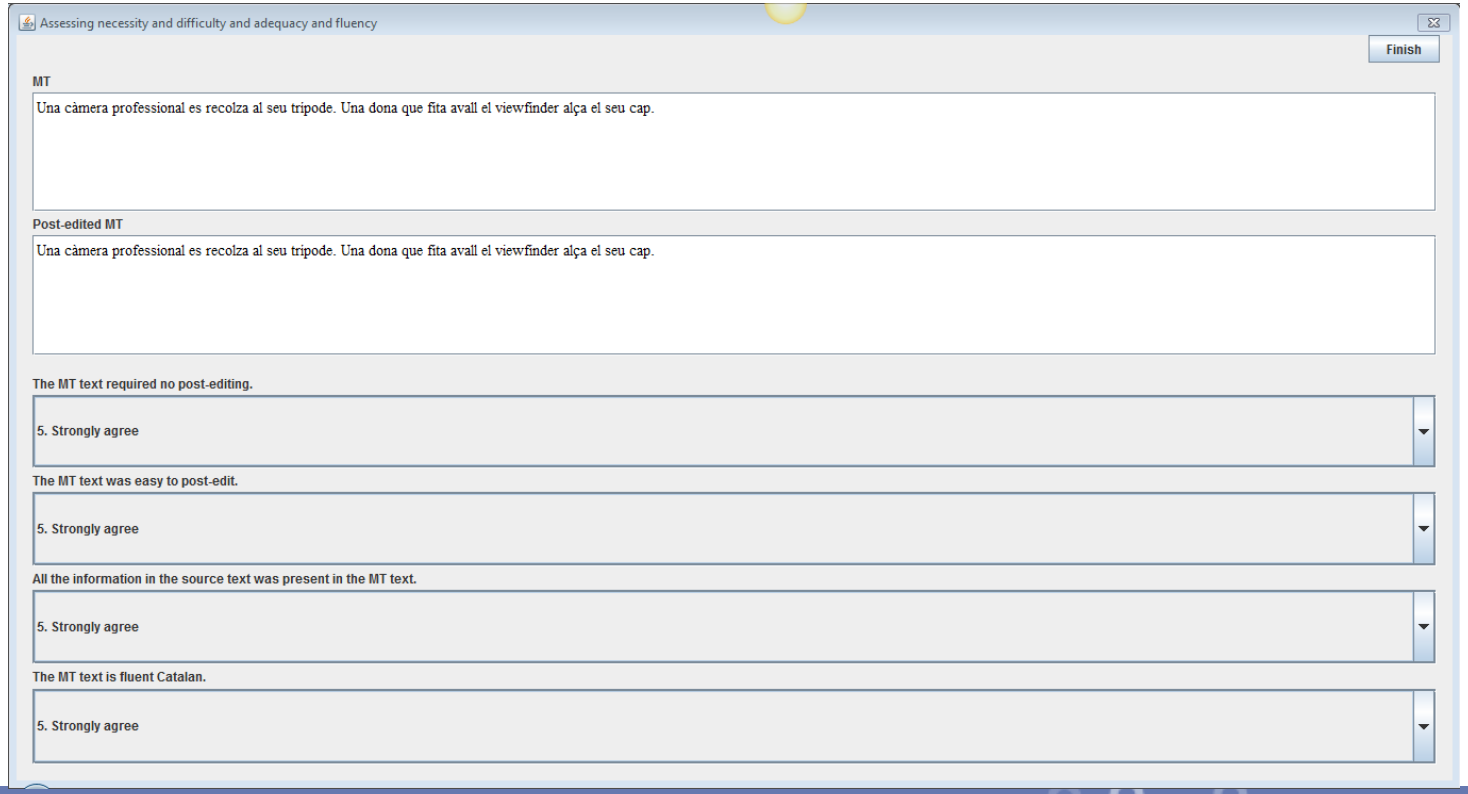
MT in AD: the pretest

- Start post-editing on PET tool

Possibility to check online resources and watch the video

MT in AD: the pretest

- Postedit unit n + questionnaire (PE difficulty & necessity, MT adequacy & fluency)



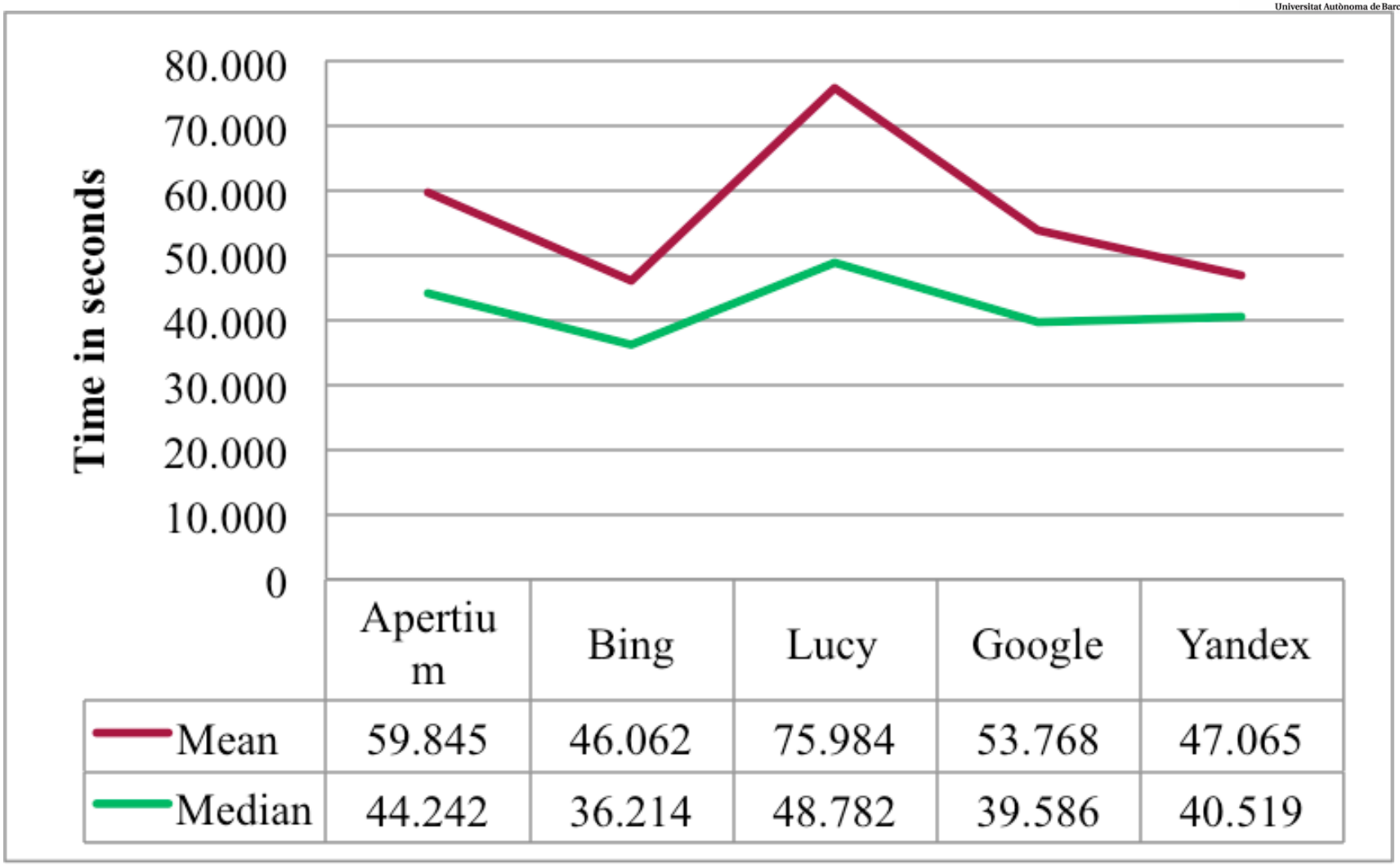
MT in AD: the pretest

- Final task: ranking (1 to 5)

	Rank the translation from best to worst, assigning numbers to each unit from 5 (best) to 1 (worst) in the left column.
	A professional camera rests on its tripod. A woman peering down through the viewfinder lifts her head.
	Unes restes de càmera professionals en el seu tripod. Una dona que mira atentament avall a través del viewfinder ascensors el seu cap.
	Una càmera professional es basa en el seu trípode. Una dona mirant cap avall a través del visor aixeca el cap.
	Un professional de la càmera es basa en el seu trípode . Una dona mirant cap avall a través del visor aixeca el cap .
	Una càmera professional es basa en el trípode. Una dona mirant cap avall a través del visor aixeca el seu cap.
	Una càmera professional es recolza al seu trípode. Una dona que fita avall el viewfinder alça el seu cap.
	Dan sits stiffly on a stool in front of a screen. The beautiful photographer turns away.
	Dan rígid asseu en un tamboret davant d'una pantalla. El fotògraf bella allunya.
	Dan està assegut rígidament en una cadira davant d'una pantalla . El fotògraf bella s'allunya .
	Dan seu stiffly en un stool davant d'una pantalla. Les voltes de fotògraf boniques fora.
	Dan seu rígidament en un tamboret davant d'una pantalla. El fotògraf bonic s'allunya.
	Dan es troba stiffly en un tamboret davant d'una pantalla. La bella fotògraf es converteix distància.

MT in AD: the pretest

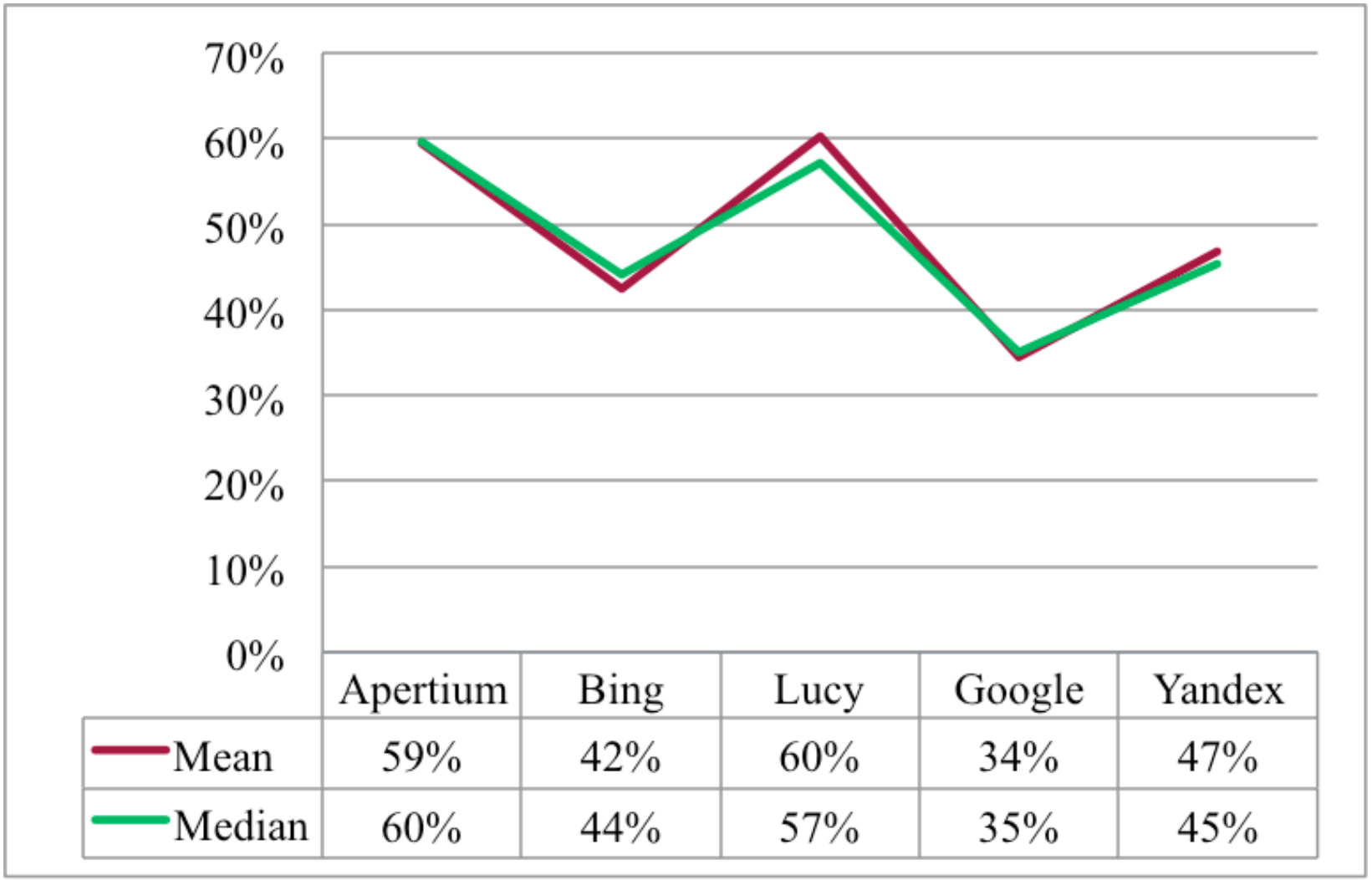
- Results: PE time
  - Bing produces the translations that require less time to be postedited, but no statistically significant differences



	Apertium	Bing	Lucy	Google	Yandex
Mean	59.845	46.062	75.984	53.768	47.065
Median	44.242	36.214	48.782	39.586	40.519

MT in AD: the pretest

- Results: HTER (median).
  - Google (35) better than others although not statistically better than Bing (44).



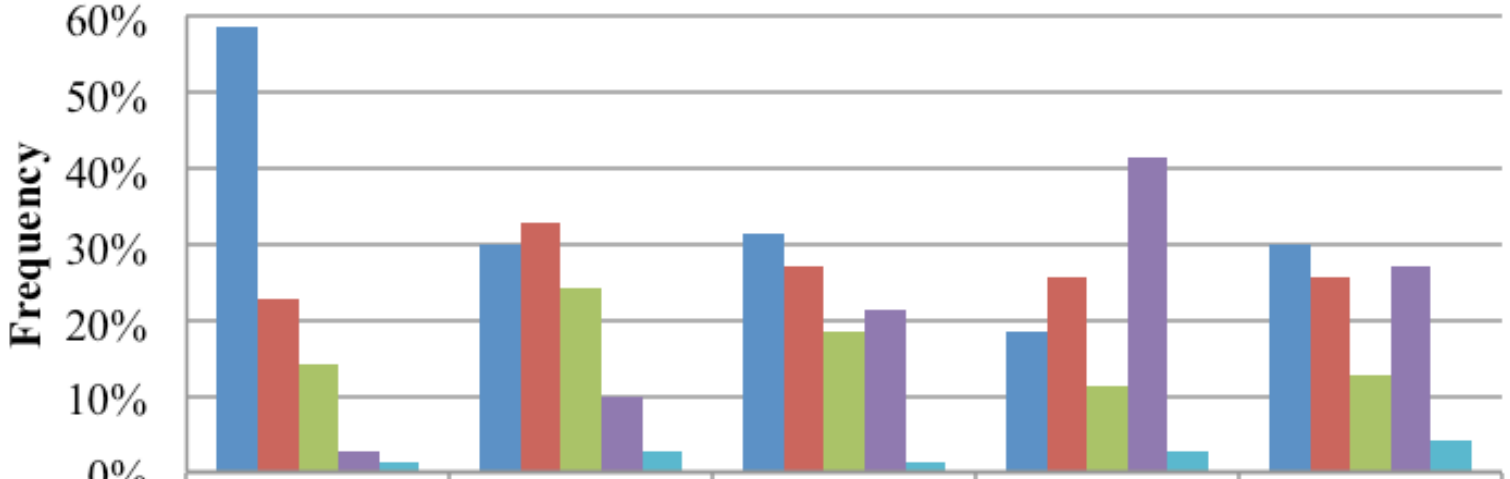
	Apertium	Bing	Lucy	Google	Yandex
Mean	59%	42%	60%	34%	47%
Median	60%	44%	57%	35%	45%

MT in AD: the pretest

- Results: PE necessity

"The MT text required no postediting"

Google obtains the highest scores (statistically significant differences): 4 or 5 in 32 out of 70 instances (44%).



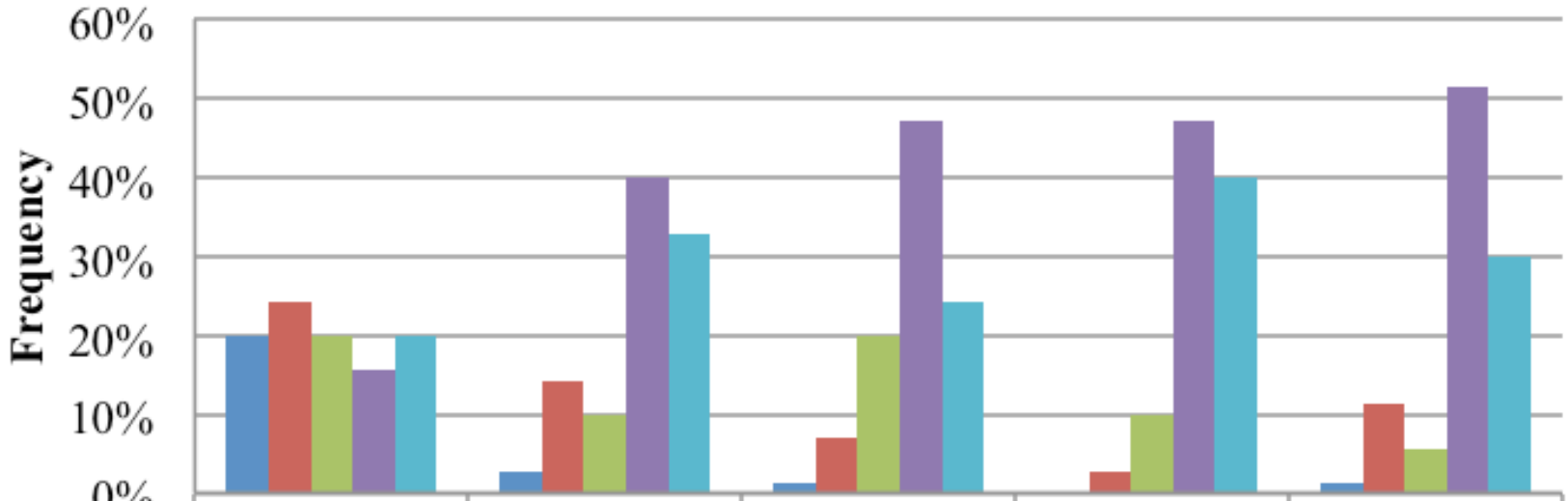
	Apertium	Bing	Lucy	Google	Yandex
1	59%	30%	31%	19%	30%
2	23%	33%	27%	26%	26%
3	14%	24%	19%	11%	13%
4	3%	10%	21%	41%	27%
5	1%	3%	1%	3%	4%

MT in AD: the pretest

- Results: PE difficulty

"The MT text was easy to postedit"

Google obtains the highest scores: 4 or 5 in 61 out of 70 instances (87%).



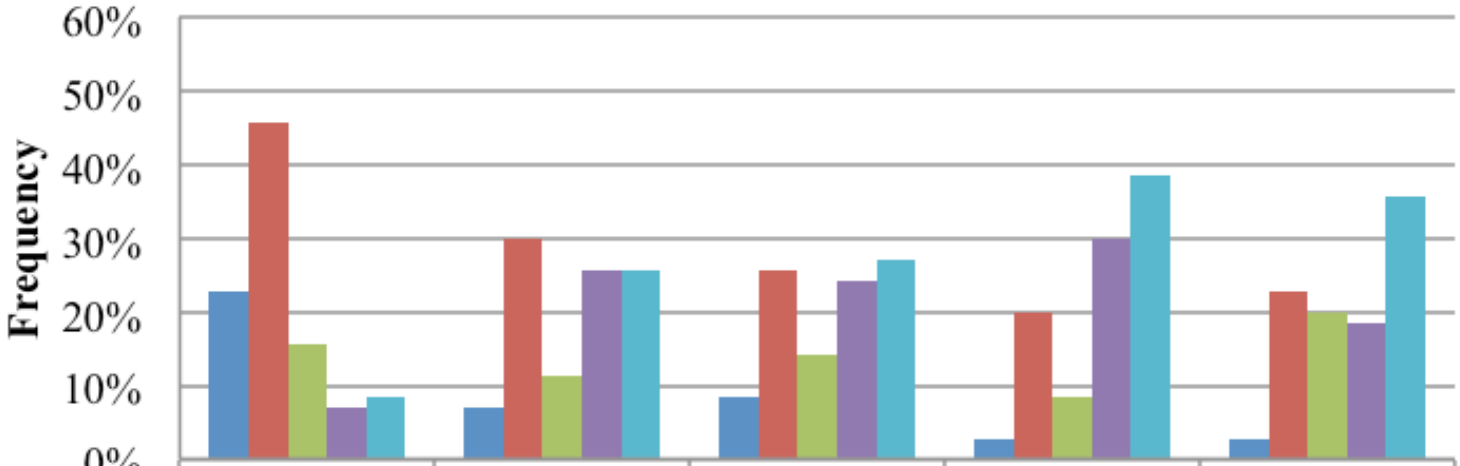
	Apertium	Bing	Lucy	Google	Yandex
1	20%	3%	1%	0%	1%
2	24%	14%	7%	3%	11%
3	20%	10%	20%	10%	6%
4	16%	40%	47%	47%	51%
5	20%	33%	24%	40%	30%

MT in AD: the pretest

- Results: MT adequacy

"All the information in the source text was present in the MT text"

Google presents higher scores (statistically significant): 4 and 5 in 69% of the instances.



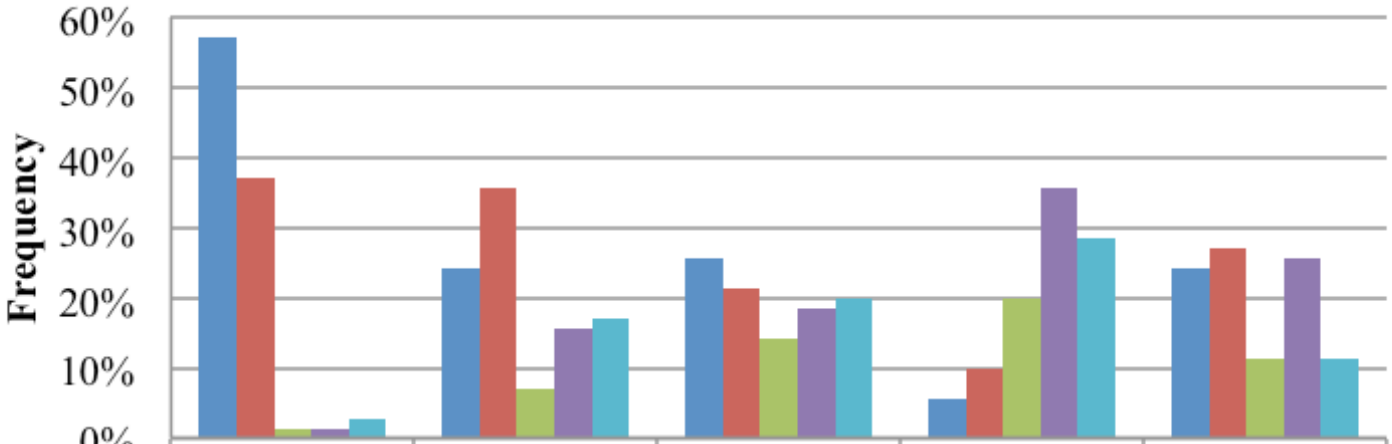
	Apertium	Bing	Lucy	Google	Yandex
1	23%	7%	9%	3%	3%
2	46%	30%	26%	20%	23%
3	16%	11%	14%	9%	20%
4	7%	26%	24%	30%	19%
5	9%	26%	27%	39%	36%

MT in AD: the pretest

- Results: MT fluency

"The Mt text is fluent in Catalan"

Google, highest scores: 55 out of 70 (65%) score 4 or 5, and only 16% score 1-2

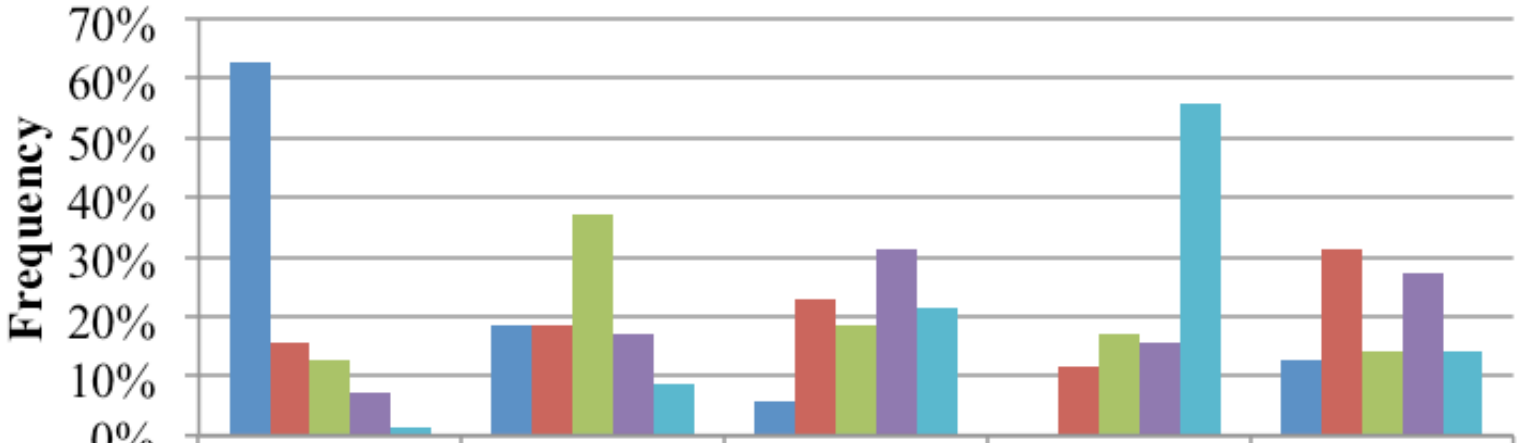


	Apertium	Bing	Lucy	Google	Yandex
1	57%	24%	26%	6%	24%
2	37%	36%	21%	10%	27%
3	1%	7%	14%	20%	11%
4	1%	16%	19%	36%	26%
5	3%	17%	20%	29%	11%

MT in AD: the pretest

- Results: ranking (from 5, best, to 1, worst)

- Google: 56% raw MT outputs ranked best, none ranked worst.



	Apertium	Bing	Lucy	Google	Yandex
1	63%	19%	6%	0%	13%
2	16%	19%	23%	11%	31%
3	13%	37%	19%	17%	14%
4	7%	17%	31%	16%	27%
5	1%	9%	21%	56%	14%

MT in AD: the pretest

- Result: Google Translate was selected for the main experiment (better scores in most evaluations)

MT in AD: setbacks

- Number of participants
- Fatigue and repetition
- Instructions (synchronisation)

… but better than subjective selection by researcher.

Ongoing work

- Ongoing analysis (finished by the end of 2014) of the comparison of creation, human translation and post-editing.

- Incorporating other technologies in the process of AD

Departament de Traducció
i d'Interpretació i d'Estudis
de l'Àsia Oriental
UAB
Universitat Autònoma de Barcelona

Machine translation and audio description: is it worth it? Assessing post-editing effort

Anna Fernández-Torné & Anna Matamala

Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental/ CAIAC, UAB/ TransMedia Catalonia

anna.torne@gmail.com
anna.matamala@uab.cat

Languages and the Media Conference 2014

Berlin, 5-7 November 2014

Funded by FFI2012-31024 (Spanish Ministerio de Economía y Competitividad) & 2014SGR27 (Catalan Government)