New tools for subtitle translation

Yota Georgakopoulou – Deluxe Media Europe, UK
Lindsay Bywood – VSI & UCL, UK
Thierry Etchegoyhen – Vicomtech-IK4, Spain

sumat

1990s: *Translation as luxury*

TM and terminology software

2000s: *Translation as commodity*

Workflow software

2010s: *Translation as utility*

Machine translation

SUBTITLING TRANSLATION INDUSTRY

YES…

- Client approved term/name lists & style guides
- Consistency issues
- Repurposing of material
- Explosion in content volumes
- Price pressure from clients
- Tighter turnaround times
- Increased use of subtitling for non-entertainment material
- Available source texts in audio language in subtitle format

SUBTITLING TRANSLATION INDUSTRY

NO…

- CAT tools
- Term banks
- Translator's workbenches
- Sufficient professional translator resources
- Increase in client budgets (despite extra volume of work)

RATIONALE

- The translation industry is increasingly embracing post-edited machine translation (PEMT). Success stories:
  - Chrysler LLC: Auto owner manuals
  - Best Western: Hotels website
  - Sony Europe: Marketing materials and catalogues
  - Sybase, a SAP company: Technical publications
- Why not test post-editing for subtitle translation?
  - Previous attempts: MUSA, eTITLE

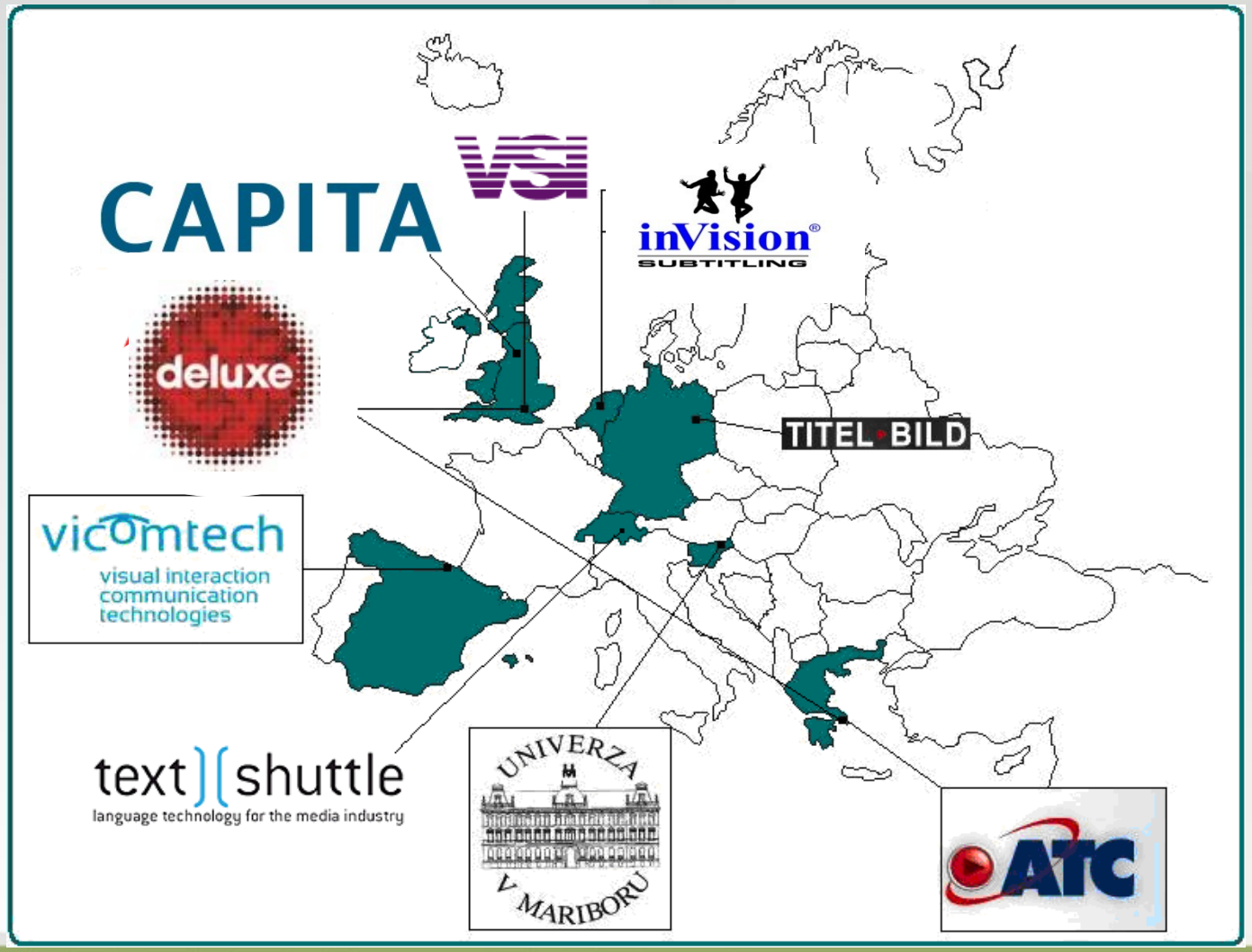
SUMAT

An Online Service for SUbtitling by MAchine Translation

http://www.sumat-project.eu

Project execution: From 01/04/2011 to 31/03/2014

sumat


CAPITA
VSI
inVision
SUBTITLING
deluxe
TITEL BILD
vicomtech
visual interaction
communication
technologies
text shuttle
language technology for the media industry
UNIVERZA V MARIBORU
ATC

sumat

1

UPLOAD SUBTITLES

2

SELECT LANGUAGES

FROM

English

TO

Dutch
French
German
Spanish
Swedish
Portugu
ese

Click to edit Master subti style

3

DOWNLOAD & POST-EDIT

MACHINE TRANSLATION

LANGUAGE PAIRS

German

French ↔ Portuguese

English

Dutch ↔ Spanish

Swedish

Serbian ↔ Slovenian

PROJECT PROGRESS SO FAR

MT SYSTEMS

- ✓ Data harvesting:
  - ✓ 7 million parallel subtitles
  - ✓ 15.5 million monolingual subtitles
- ✓ Data processing:
  - Classified according to genre
  - Automatically aligned
- ✓ SMT engine building:
  - ✓ Experiments with linguistic rules
  - ✓ Additional open source data used
  - ✓ Various SMT models built & mixed

ONLINE SERVICE

- ✓ Live demo
  http://online.sumat-project.eu/sumat/web/guest/live-demo
- ✓ Online service prototype

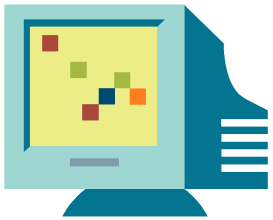
…TO BE EVALUATED!

SUMAT EVALUATION

CASE STUDY: JULY 2012 – OCTOBER 2012 (COMPLETE)
1st ROUND (3 PHASES): APRIL 2013 – SEPTEMBER 2013 (COMPLETE)
2nd ROUND: OCTOBER 2013 – FEBRUARY 2014

Human evaluation
Quality scoring
Error classification
Subjective evaluation
Timed Post-Editing

Automatic evaluation
BLEU
METEOR
TER
Equal
Levenshtein 5



EVALUATION ROUNDS

Round 1

- Measure MT quality through human ranking
- Provide general feedback on the post-editing experience
- Collect recurrent errors
- Improve quality of SMT systems

Round 2

- Measure productivity gain/loss through post-editing
- Evaluate final SMT systems in a professional use-case scenario

Round 1: Design (I)

- Translation pairs:
  - EN into DE, ES, FR, NL, PT, SV
  - ES, FR, DE into EN
  - SL < - > SR

- Adapt SMT systems after each post-editing phase

Phase 1	Phase 2	Phase 3
April	June	August
Post-editing	Post-editing	Post-editing
2 input text files	2 input text files	1 input text file
2 video files	2 video files	1 video file
4 MT output files	4 MT output files	2 MT output files

Round 1: Design (II)

Subtitlers were asked to:

- Post-edit to their usual quality standards
- Score each individual subtitle on a 1 (bad) to 5 (good) scale
- Mark recurrent errors according to a supplied taxonomy for subtitles ranking 3 or higher
- Fill in a questionnaire about their experiences and give opinions on the MT output

Round 1: Quality scale

- **1**: The machine translated subtitle is incomprehensible and requires a new translation from scratch.
- **2**: About 50% to 75% of the machine translated subtitle needs to be edited.
  It requires a significant editing effort in order to reach publishable level.
- **3**: About 25 to 50% of the machine translated subtitle needs to be edited.
  It contains various errors and mistranslations that need to be corrected.
- **4**: About 10 to 25% of the machine translated subtitle needs to be edited.
  It is generally clear and intelligible.
- **5**: The machine translated subtitle is perfectly clear and intelligible.
  It is not necessarily a perfect translation, but requires little to no editing.

Round 1: Error taxonomy

- **Agr**: Any kind of agreement error (subject-verb, article-noun, etc.).
- **Miss**(ing): Any translation where part of the original subtitle is missing.
- **Order**: Any translation with incorrect word order.
- **Phrase**: Any group of words that should have been treated as a unit but were translated separately, or any group of words that were treated as a unit but should have been translated separately.
- **Cap**: Any word which should be either lower-cased or upper-cased.
- **Punc**: Any missing or spurious punctuation.
- **Spell**: Any misspelled word.
- **Length:** Subtitles that are too long.
- **Trans**: Poor or wrong choice of word translation, or word left in original language.

Round 1: Material

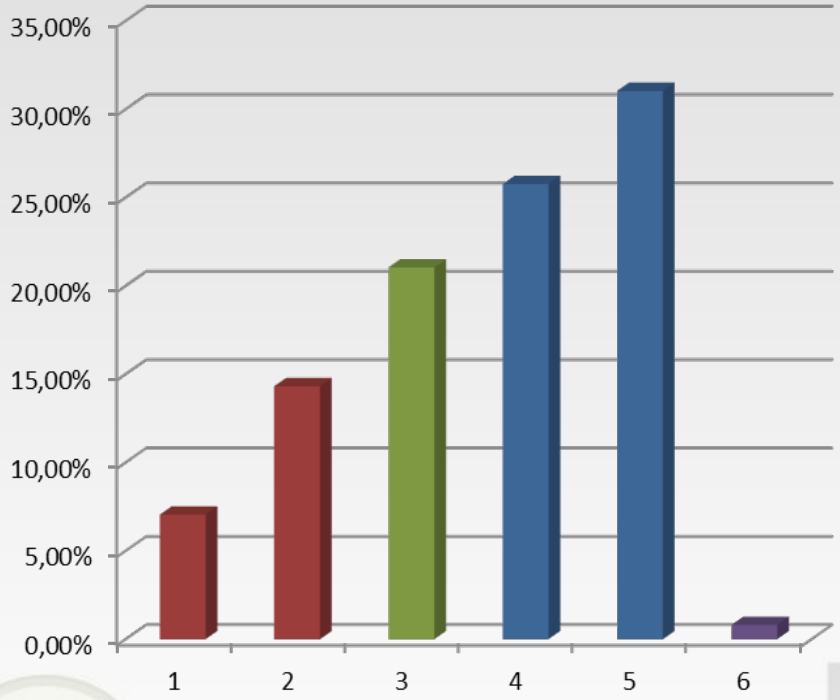
- Various types of input files:
  - Scripted & unscripted
  - Different domains/genres (e.g. drama, documentaries, magazine programmes, corporate talk shows)
- Input files not used for training/tuning the systems
- Total of 27 565 post-edited, ranked & annotated subtitles
  - Phase 1:13 602
  - Phase 2: 10 643
  - Phase 3: 3 320
- Post-editing performed with subtitling software of choice

Round 1: Evaluation goals

- Evaluate the quality output of the main combinations of MT systems
- Apply automated metrics to post-edited files
- Measure variation per translation pair
- Measure correlation between human ranking and automated metrics

  - ✧ Compare several systems combinations:
    - ✧ Professional vs. crowd-sourced corpora
    - ✧ Different domains: open (SUMAT, OpenSubs), European (Europarl, EuroparlTV), scientific (TED)
    - ✧ Results shown for optimal MT system: SUMAT+OpenSubs+Europarl

Round 1 results: Ranking

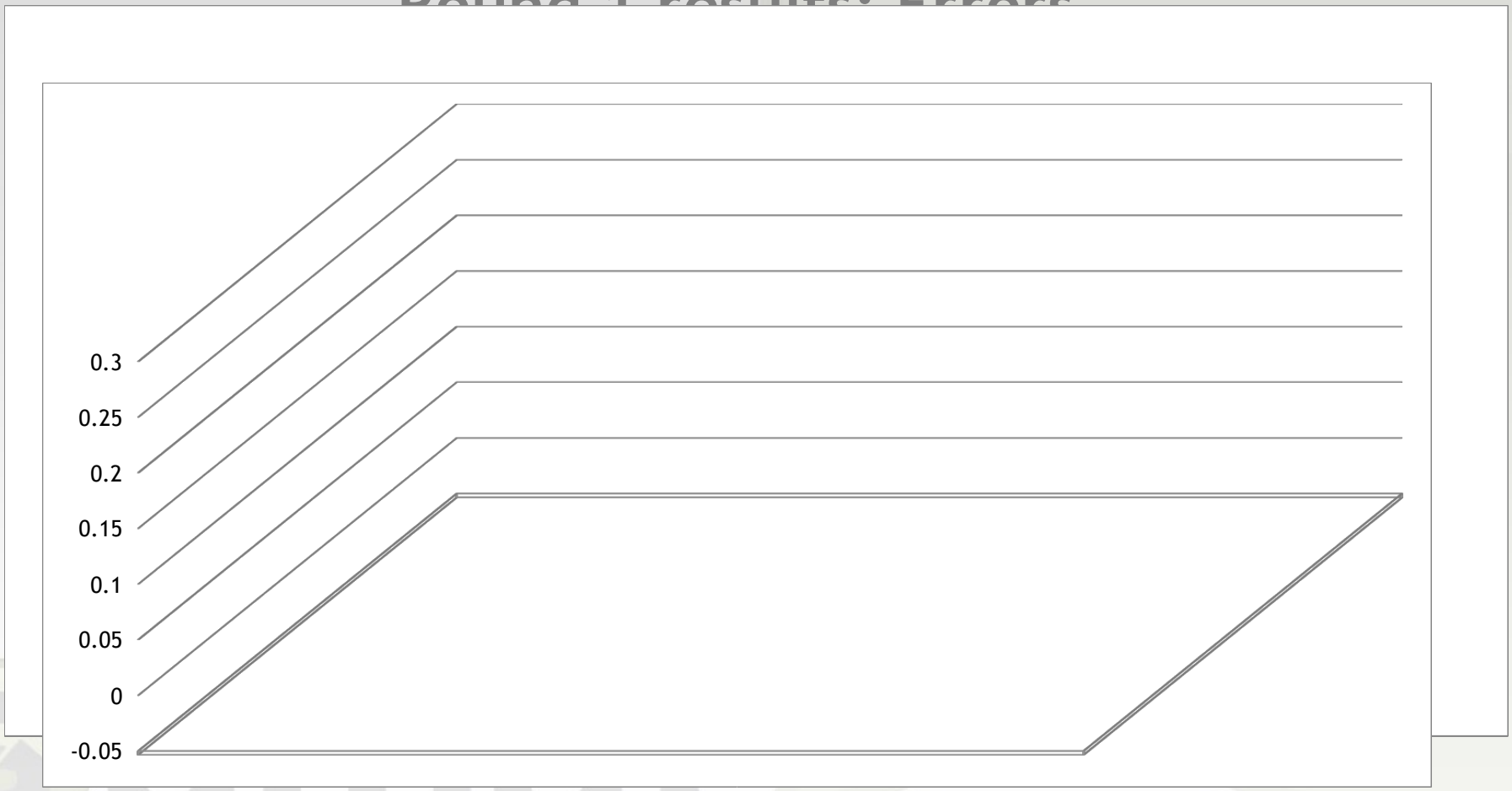


sumat

Round 1 results: Metrics

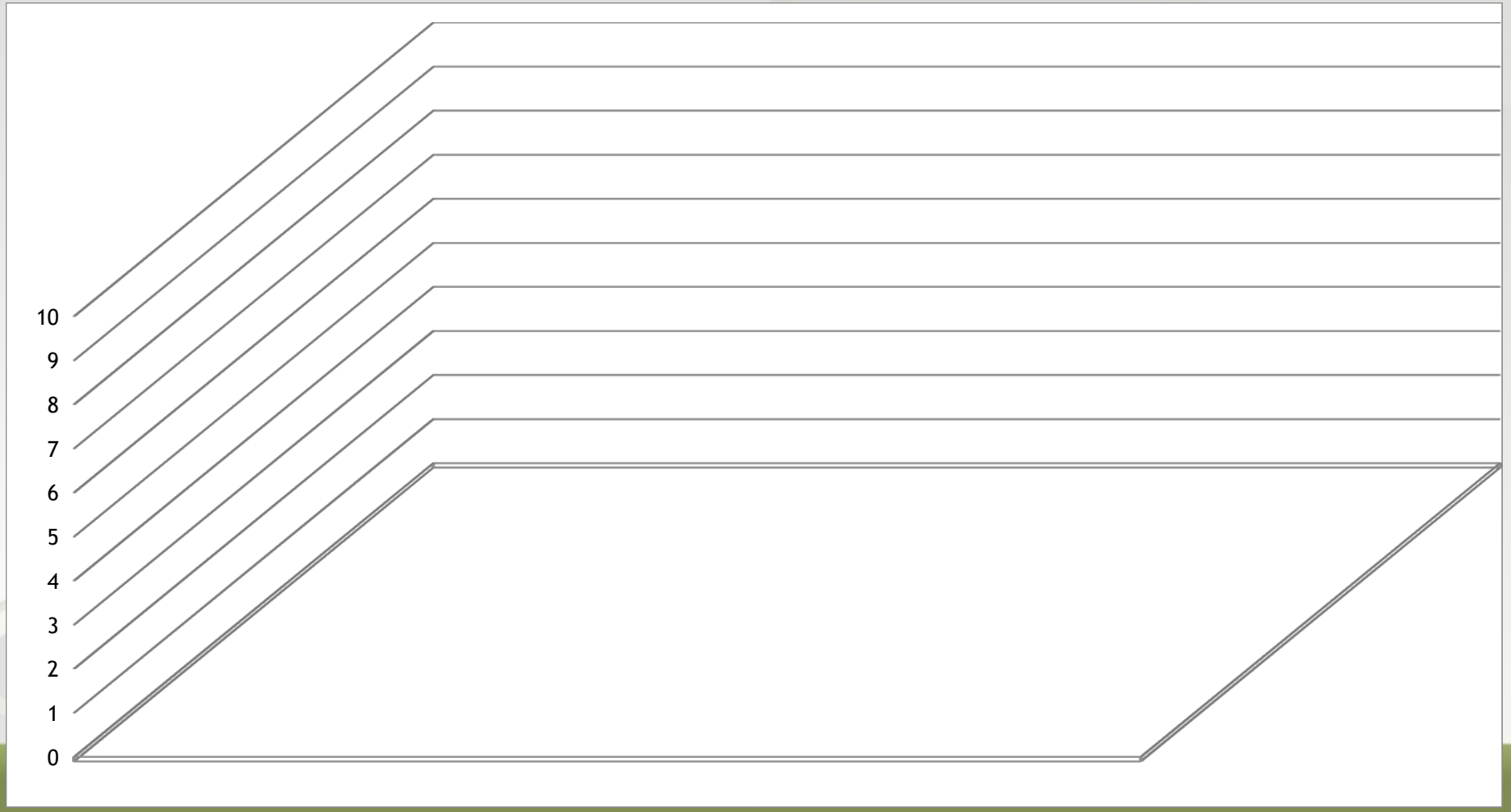
sumat

Round 1 results: Errors



sumat

Language pairs comparison



sumat

Examples

Source text		MT output	
EN	-Bill? -I don't understand what's gotten into you.	ES	- ¿Bill? - No entiendo qué te ha picado.
EN	How long are you gonna give her a free pass?	ES	¿Cuánto tiempo le vas a dar vía libre?
EN	You still don't understand, do you?	DE	Du verstehst es immer noch nicht, oder?
EN	What are you still doing here?	DE	Was machst du denn noch hier?
EN	Honey, I'm a little worried about your mother.	DE	Schatz, ich mache mir etwas Sorgen um deine Mutter.

There were many fixed phrases that were correct and usable.

My perception is that there is still a long way to go, however I must admit that some of the translations given impressed me.

With shorter and simpler sentences like the ones in this episode, I think having the translation there saves quite some time.

Examples

Once I got it going, it was quite easy.

Source text		MT output	
EN	What do you have to do to get a drink around here?	FR	Que doit-on faire pour avoir un verre ?
EN	- Can you do it? - I'll have a go.	FR	- Tu peux le faire ? - Je vais essayer.
EN	Where we at?	FR	On en est où ?
ES	- Sí. - En ese sentido, eres especial.	EN	- Yes. - In that sense, you're special.

Some subtitles (very, very few) were very good, with options that I hadn't thought of.

In many aspects the quality is at times pretty good.

The simpler the subtitle, the better the quality of the machine translation.

PE: Perceptions & Variations

- EN2SV:
  - *"Hugely improved since last year! I have many 4 and 5 and am really quite amazed. There's still a long way to go, but it's usable already now."*
  - *"I think this type of material is too complex for translation tests of this kind."*
- FR2EN:
  - *"Overall pretty good. Simple sentences were usually perfect, but the machine has problems when the sentence is complicated […]"*
  - *"[…] the content quality was still quite poor […] Very few subtitles were left unchanged."*
- *EN2ES:*
  - *"[…] I guess all in all everything depends on the type of show being subtitled."*
  - *"For this kind of program, it surprised me that the quality of the machine translated subtitles was quite good even though the language used is quite colloquial".*

Round 1: Summary

- General feedback:
  - Heavy load on translators: annotating, classifying & post-editing
  - MT helps in cases of minor to moderate post-editing
  - Frustrating with bad translations & extra effort in determining what to do with MT output when checking bad or in-between cases
  - Easier to deal with MT material after some post-editing practice
  - Several evaluators surprised by MT quality/fluency when correct

- Best systems obtained by mixing translation models

- Global metrics
  - More than half (56.79%) of MT subtitles were ranked 4 or 5
  - High numbers of Equal & Lev5 – Good averages on metrics
  - Good correlation levels between human and automated evaluation

Round 2: Evaluation goals

- Measure productivity gain/loss in a commercial use-case scenario
  - Translate from source – benchmark
  - PE full MT files
  - Translate & PE filtered MT output

- Address post-editors' main frustration in Round 1 by automatically filtering out poor MT output

- Evaluate two opposite cases in subtitling:
  - Scripted files – Easier for MT
  - Unscripted files – Most difficult for MT

Findings

- Logistics
- Many variables in this type of MT evaluation: workflow, material, translator expectations, effort in assessing MT quality
  - Translators' quality scores climb consistently from poor to good
  - Good results overall in terms of volume of almost ready to use MT output in the subtitle domain
  - Need for integrated MT quality assessment => automatically filter out poor MT cases before post-editing
  - Need for a better explanation of linguistic phenomena currently out of MT reach => post-editor training

Conclusions

- PEMT: different and new way of doing subtitle translation
- Productivity boost?
- Needs to be measured exactly per language pair

- Sensitivity of translation quality to programme material
- Good correlation between human judgement and automated metrics

- Translators' expectations & perception of task
- Effort in assessing MT quality before post-editing

- Rising demand for post-editing skills in the subtitling industry

Thank you!
Questions?

Yota.Georgakopoulou@bydeluxe.com
Lindsay.Bywood.13@ucl.ac.uk
tetchegoyhen@vicomtech.org