SUMAT: Building on our experience

Lindsay Bywood - VSI and Imperial College, UK
Thierry Etchegoyhen - Vicomtech, Spain
Mark Fishel - TextShuttle, Switzerland
Yota Georgakopoulou - Deluxe Media Europe, UK
Martin Volk - University of Zurich, Switzerland

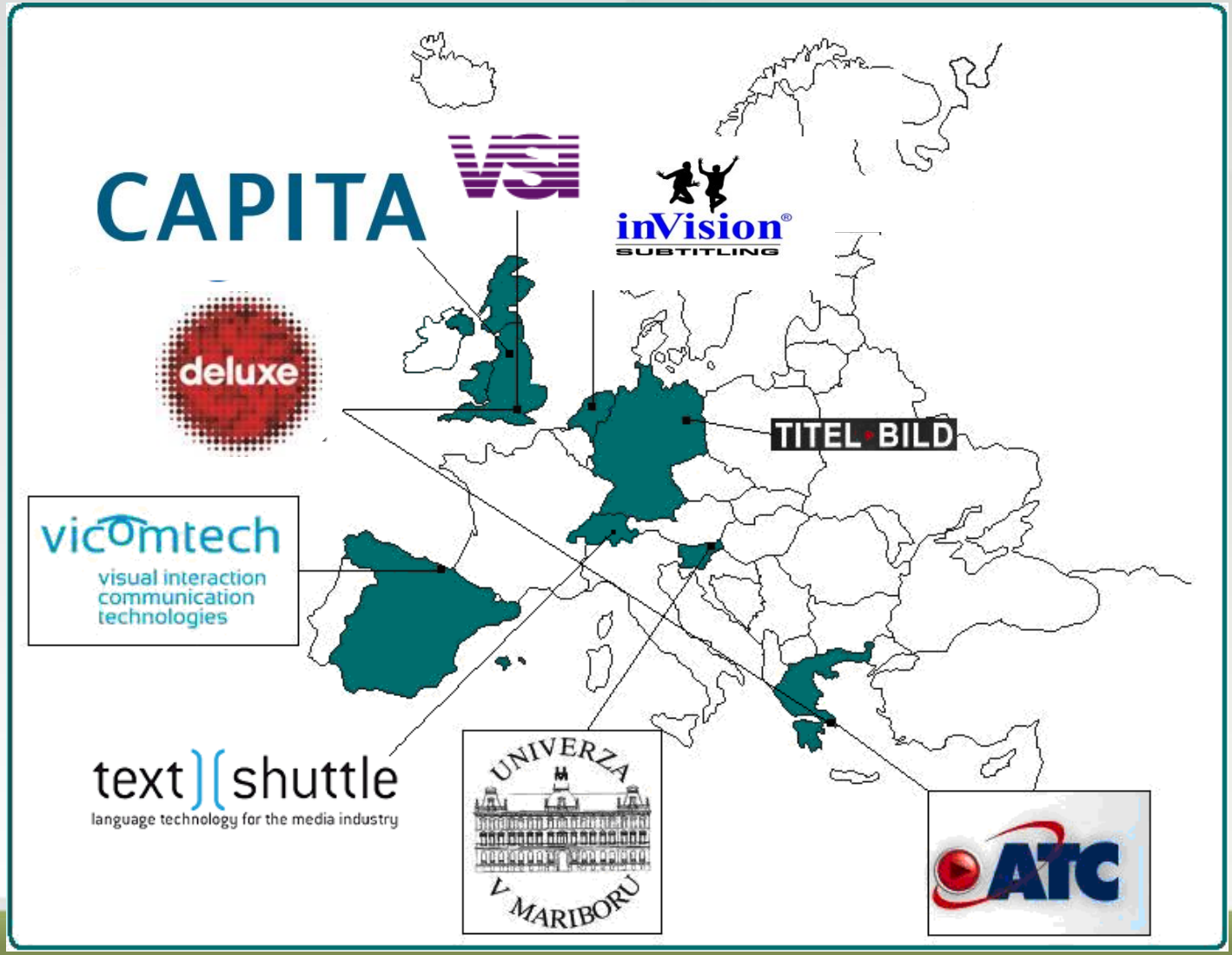
SUMAT

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

UPLOAD SUBTITLES

1

SELECT LANGUAGES

2

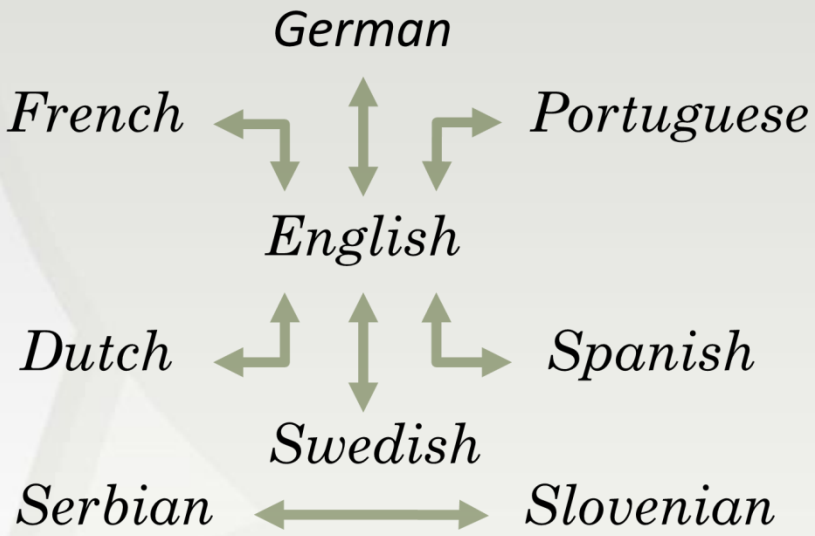
FROM	TO
English	Dutch French German **Spanish** Swedish Portuguese

DOWNLOAD & POST-EDIT

3

MACHINE TRANSLATION

LANGUAGE PAIRS

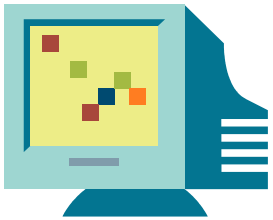


SUMAT EVALUATION

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

- Baselines:
  - Statistical machine translation systems
  - Trained on professionally created corpora provided by subtitling companies in the SUMAT consortium
  - Average of 1.1M aligned subtitles per language pair and 15.5M monolingual subtitles overall

- Augmented systems built by:
  - Selecting additional data from different domains
  - Training separate translation models
  - Testing various combinations of models

- Additional corpora experimented with:
  - OpenSubs (92 330 443 total aligned subtitles – crowd-sourced)
  - Europarl (15 600 608 total aligned sentences - pro)
  - TED (823 298 total aligned subtitles – crowd-sourced)
  - Europarl TV (991 605 total aligned subtitles - pro)

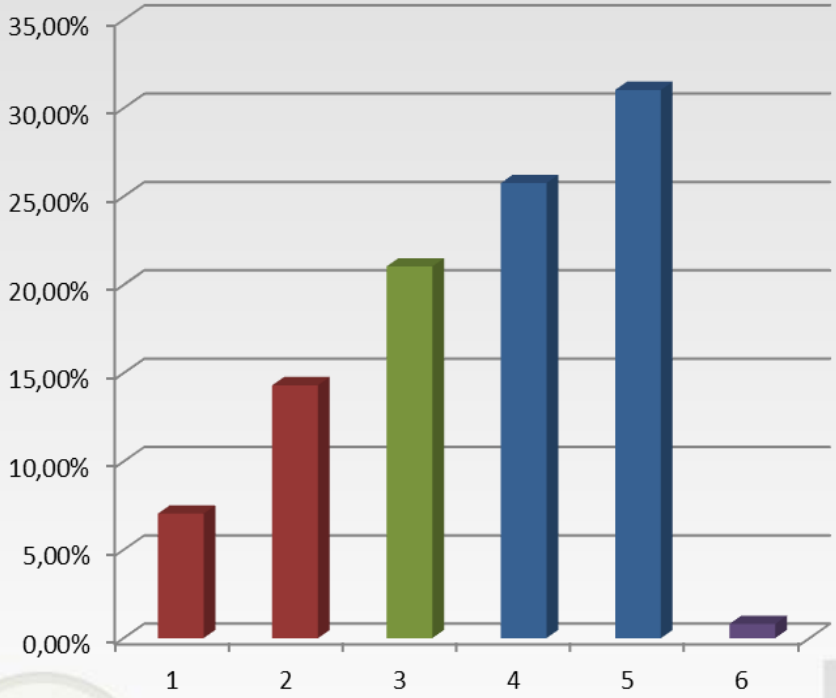
Round 1: Evaluation goals

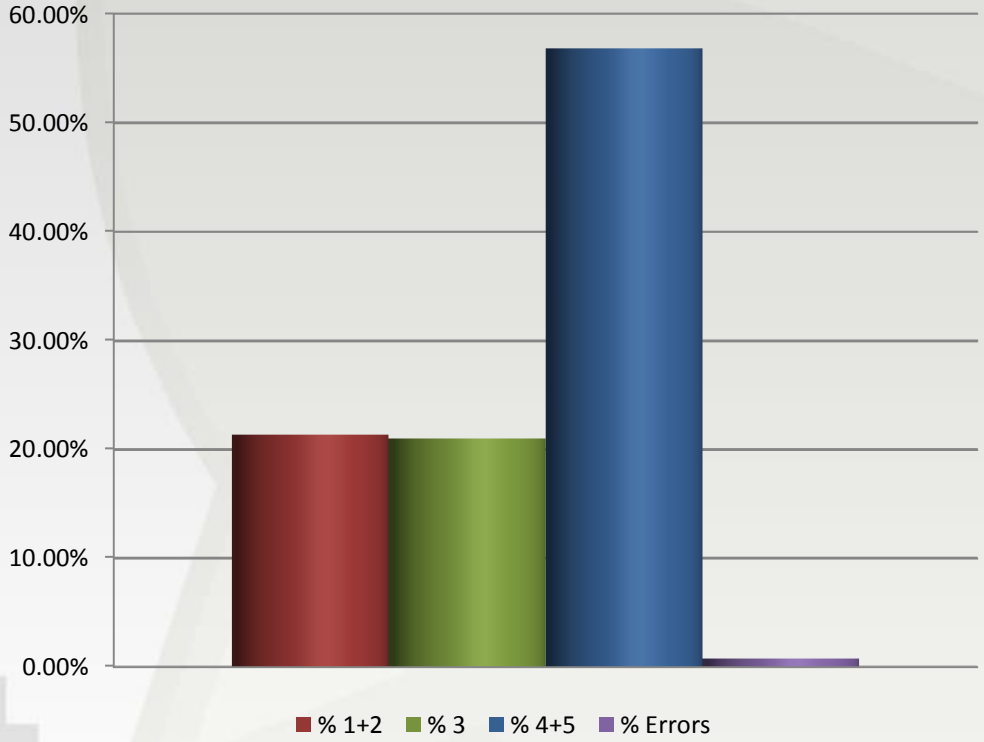
- Evaluate the quality output of the main combinations of MT systems
- Apply automated metrics to post-edited files
- Measure variation per translation pair
- Measure correlation between human ranking and automated metrics

✧ Optimal MT system: SUMAT+OpenSubs+Europarl

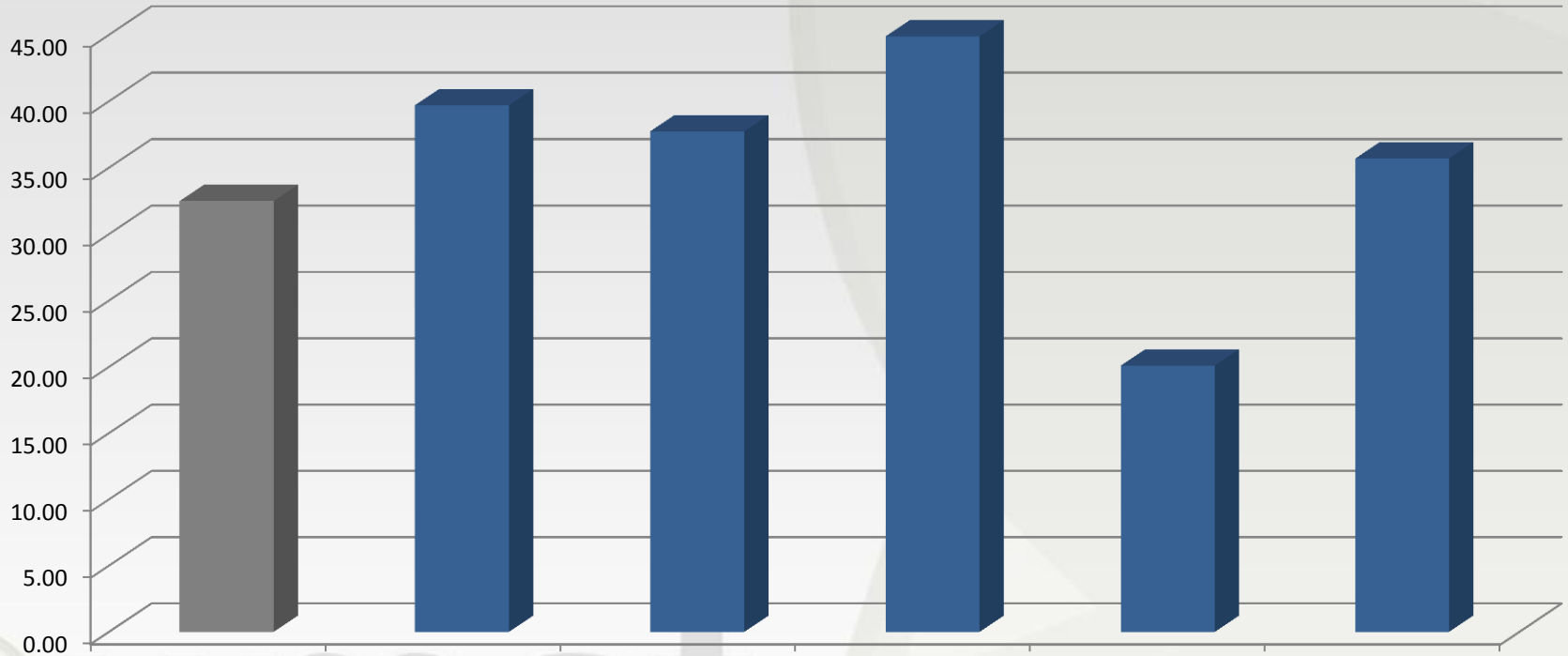
sumat

Round 1 results: Ranking





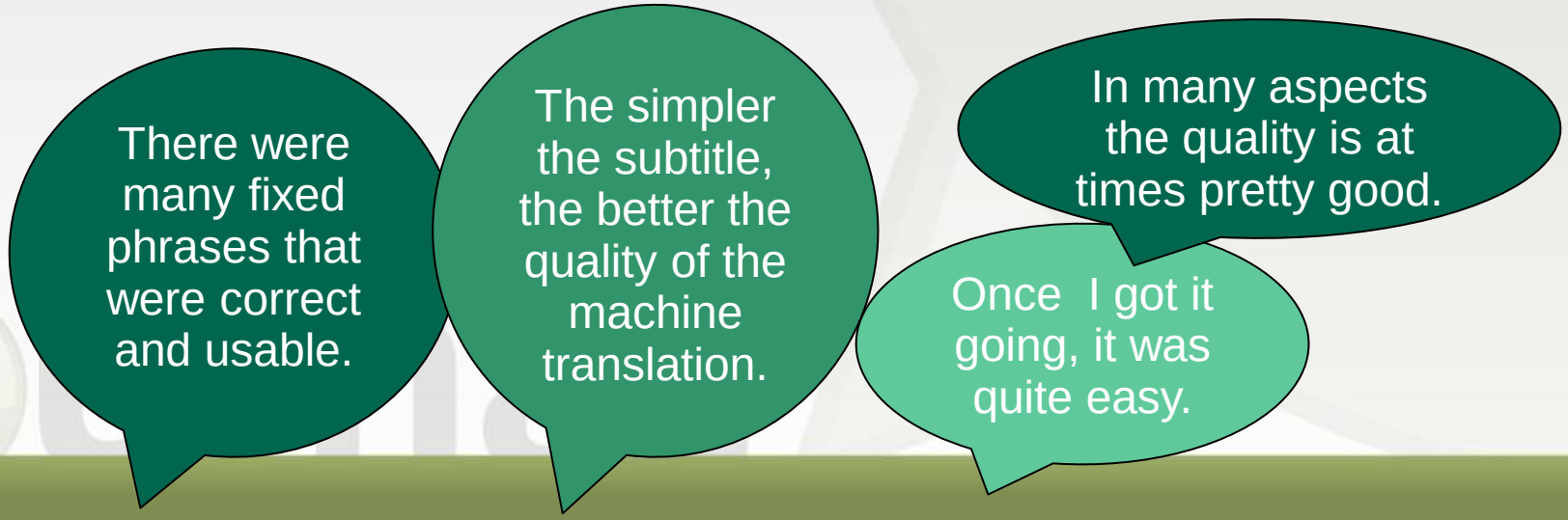
Round 1 results: Metrics



Avg ranking	BLEU	METEOR	TER	Equal	Lev5
3.60	**39.69**	**37.73**	**44.88**	**20.10**	**35.69**

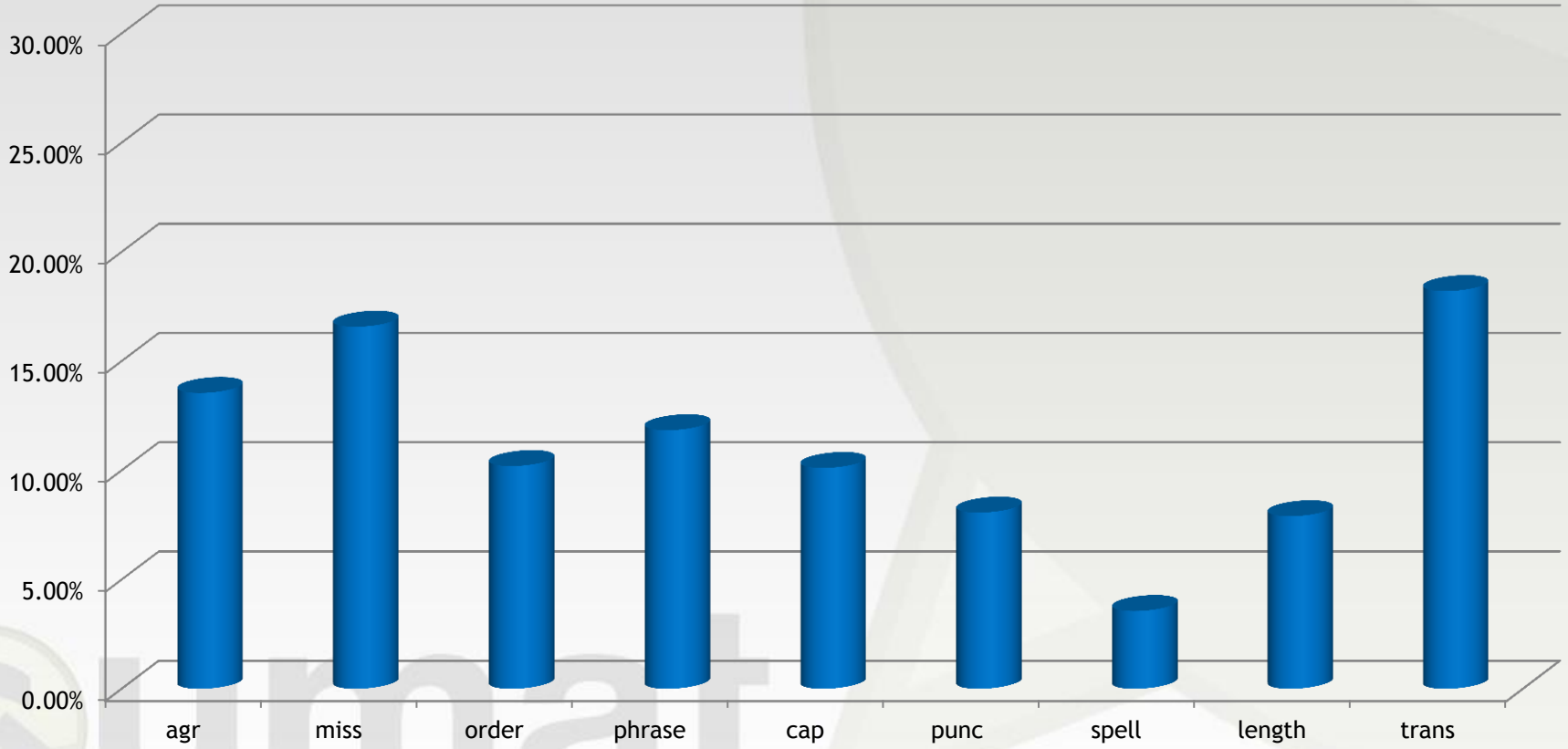
sumat

English source text	MT output
-Bill? -I don't understand what's gotten into you.	- ¿Bill? - No entiendo qué te ha picado.
How long are you gonna give her a free pass?	¿Cuánto tiempo le vas a dar vía libre?
You still don’t understand, do you?	Du verstehst es immer noch nicht, oder?
What are you still doing here?	Was machst du denn noch hier?



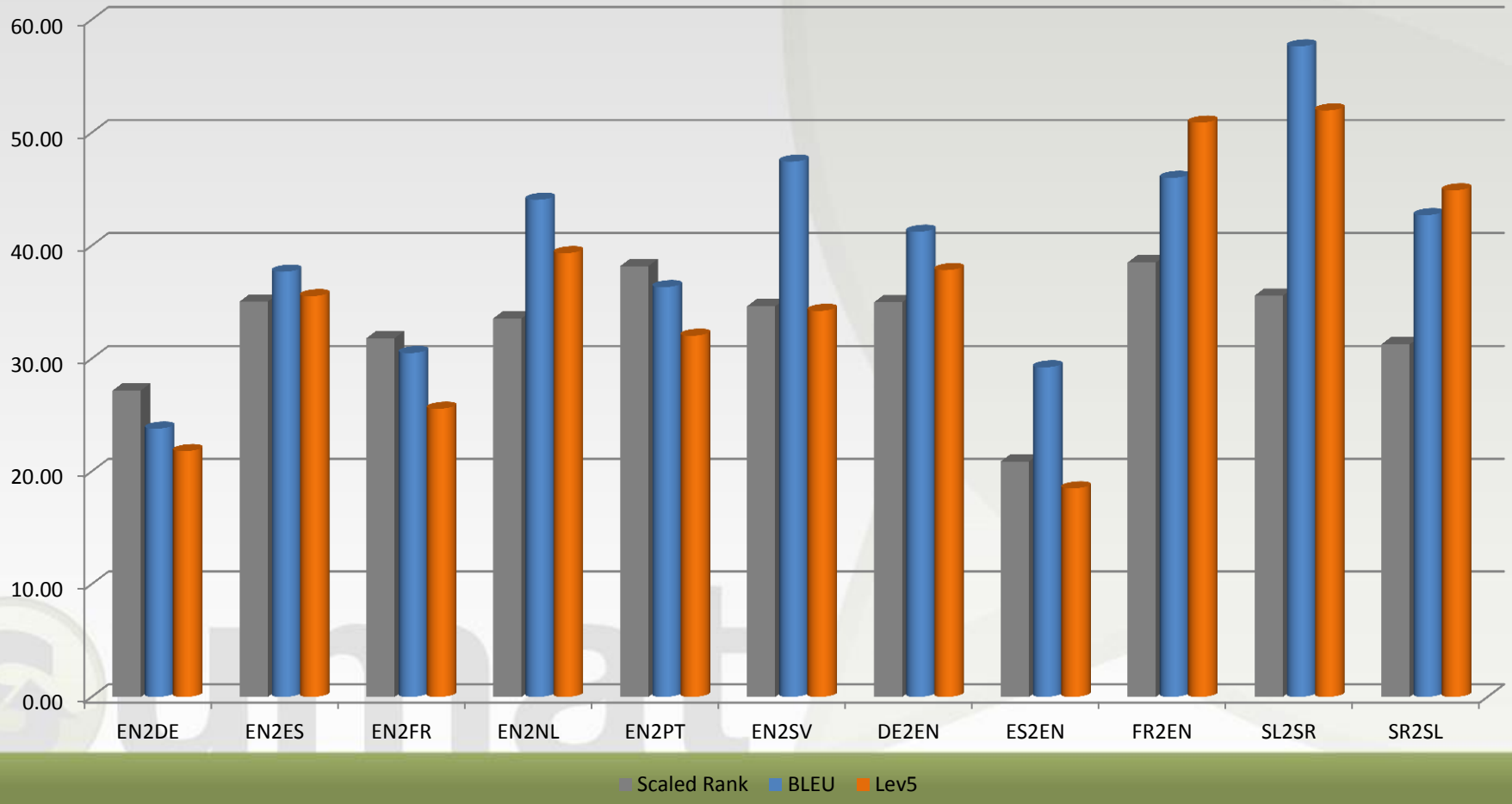
sumat

Round 1 results: Errors



sumat

Language pairs comparison



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:Design

- Measure productivity gain/loss:
  - 2 translators per translation pair
  - 4 files to be posted-edited
  - 2 benchmark files translated from scratch
  - Timed post-editing with subtitling environment of choice
  - Same translation pairs as in Round 1

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

- Appraise evaluation (SUMAT vs. System Z)

Findings & Conclusions

- 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 subtitles domain
- Need for a better explanation of linguistic phenomena currently out of MT reach
- Need for integrated MT quality assessment => automatically filter out poor MT cases before post-editing

Thank you!
Questions?