# A SYNTHESIS OF HUMAN AND MACHINE

## CORRELATING "NEW" AUTOMATIC EVALUATION METRICS WITH HUMAN ASSESSMENTS

Presenters: Andrea Alfieri, Mara Nunziatini

# AGENDA

01

OBJECTIVES AND METHOD

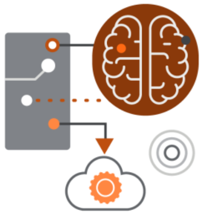
02

RESULTS

03

CONCLUSIONS AND FURTHER RESEARCH

# Objectives And Method

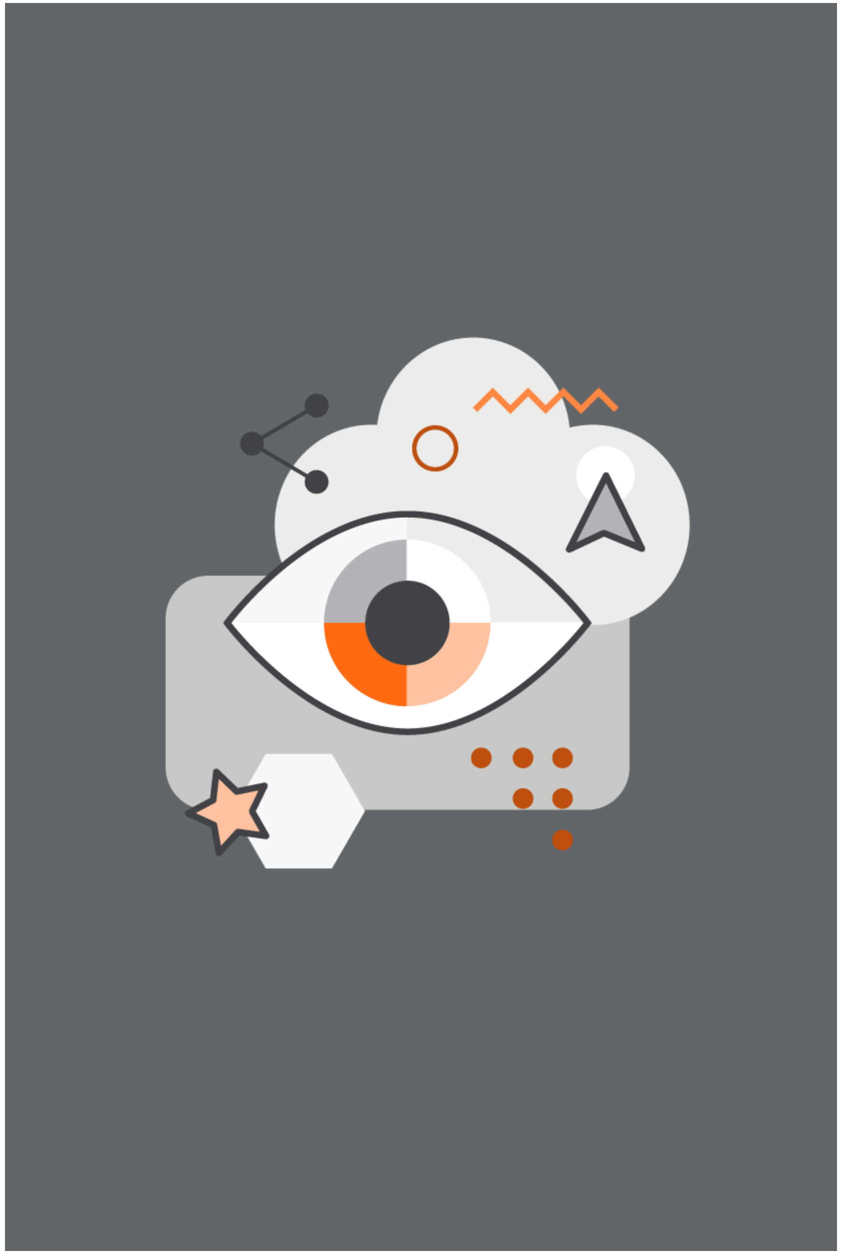


## Objectives

- Provide an overview of new Machine Translation metrics: **characTER, chrF3, COMET, hLEPOR, Laser, Prism**.
- Analyze if and how these metrics correlate at a segment level to the results of Adequacy and Fluency **Human Assessments**.
- Analyze how they compare against **TER** scores and **Levenshtein Edit Distance** as well as against each of the other.

## Method

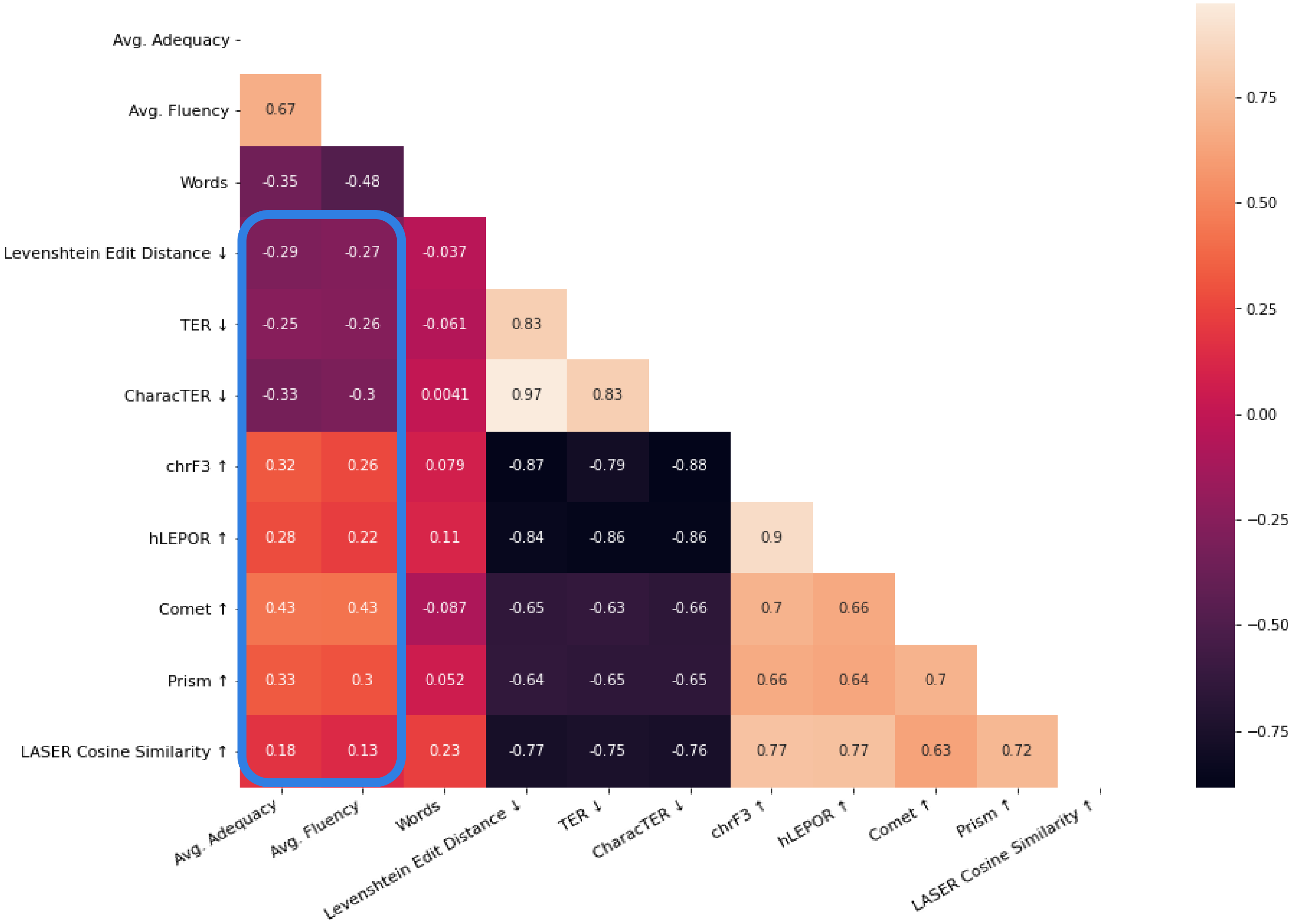
1. **~500 segments** (~ 250 UI/UA + ~ 250 Marketing) selected for the experiment and scored for Adequacy and Fluency
    - Adequacy and Fluency: scores from 1 (lowest) to 5 (highest)
    - **3** experienced **linguists** per language (scores averaged)
    - Languages: **German, Hindi** (no model for Prism)**, Italian, Russian, Simplified Chinese**
2. The same segments were scored using characTER, chrF3, COMET, hLEPOR, Laser, Prism, TER and Levenshtein Edit Distance
3. Human Assessment scores  and Automatic Scores aligned and analyzed (Pearson Correlation Coefficient)



# Results

## Pearson Correlation Coefficient per Metric and Language

# German

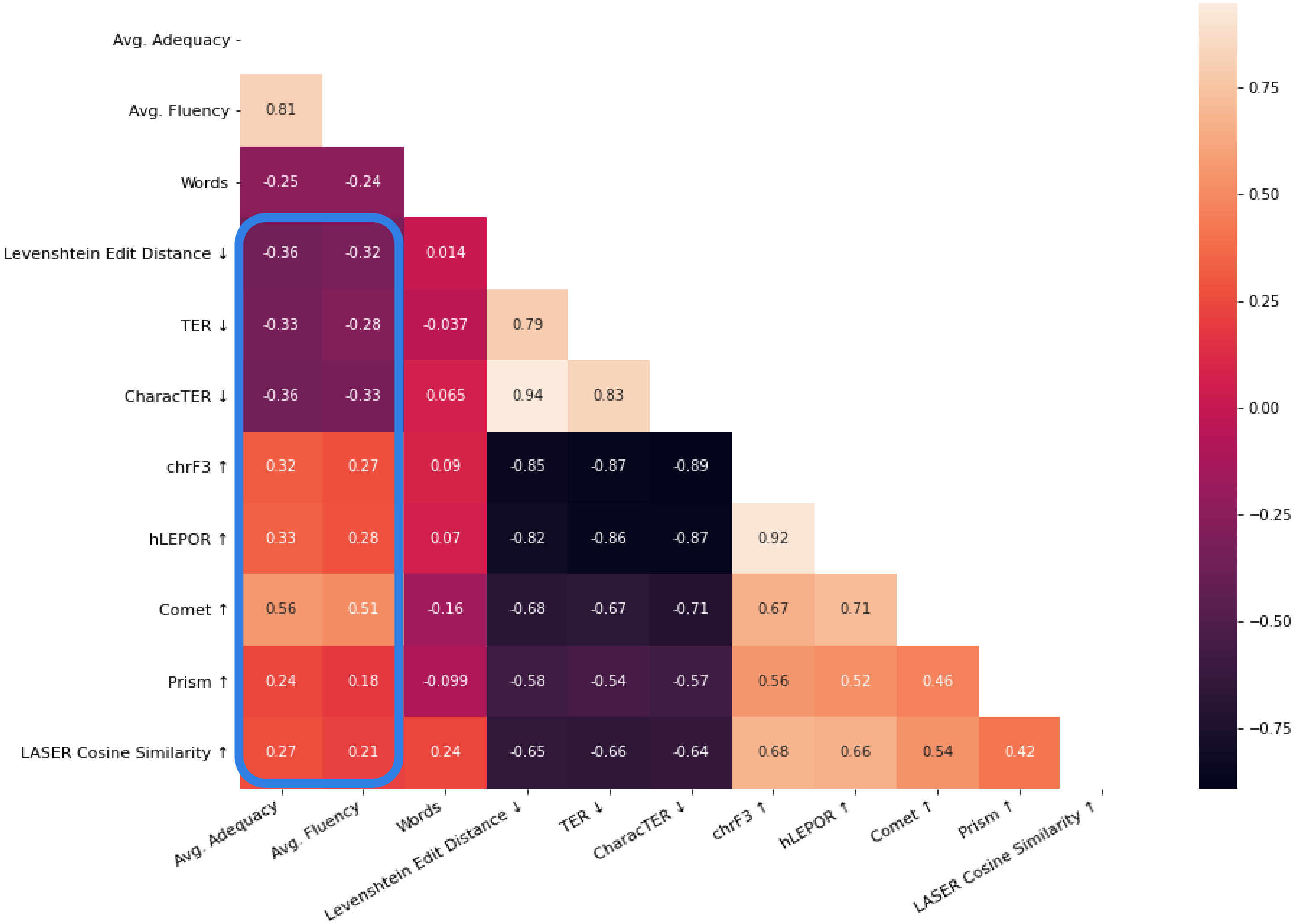



**Insights**

Pearson Correlation Coefficient calculated to analyze the correlation between Human Assessment and metrics, as well as between each one of the metrics included in the study.

- **COMET** is the metric that achieves the best correlation with Human Assessments.
- The second place goes to Prism and CharacTER, which show comparable results.
- The third place goes to chrF3.
- Levenshtein Edit Distance and TER show a worse correlation compared to the 3 new metrics mentioned above.

# Hindi

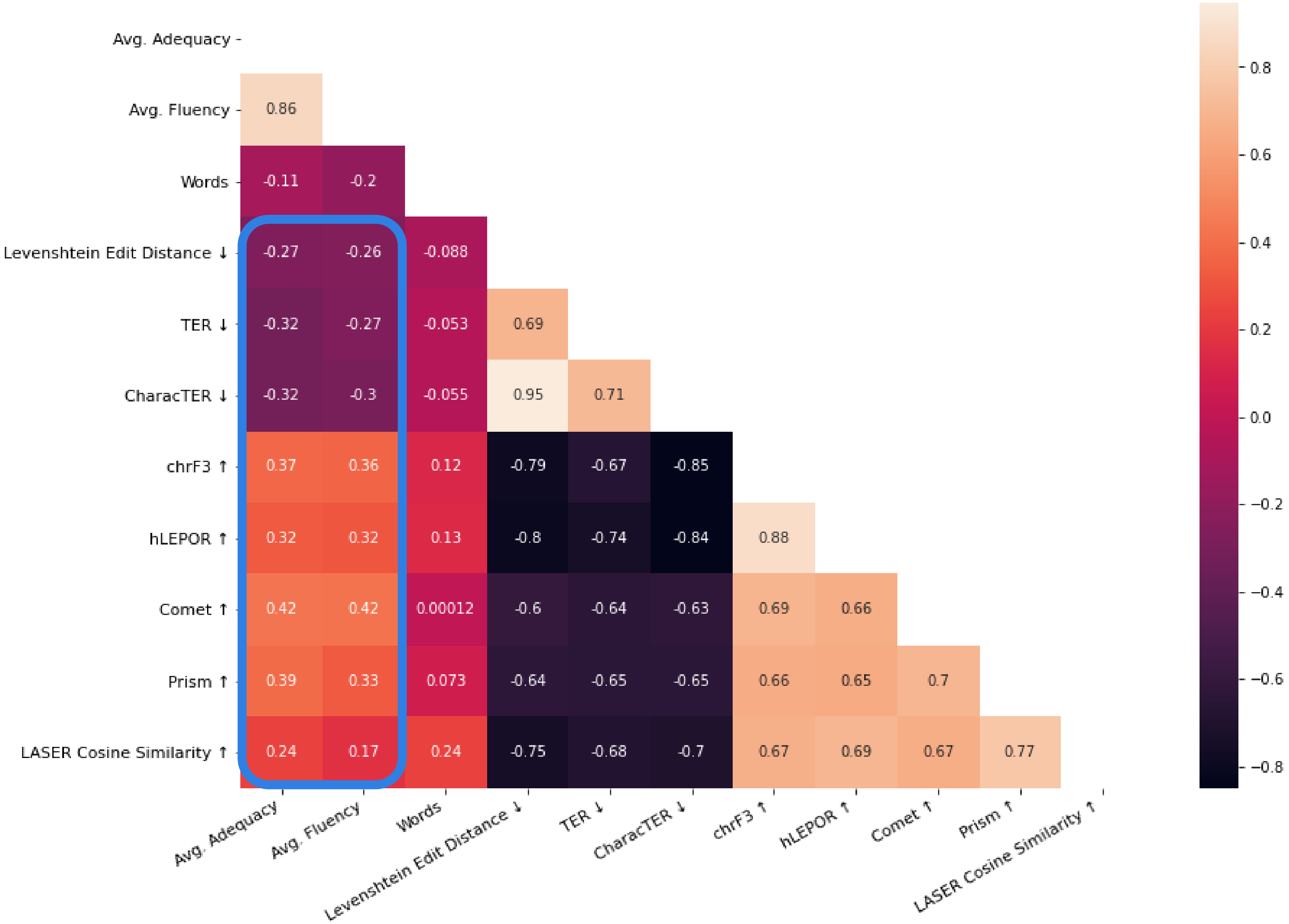



### Insights

Pearson Correlation Coefficient calculated to analyze the correlation between Human Assessment and metrics, as well as between each one of the metrics included in the study.

- **COMET** is the metric that achieves the best correlation with Human Assessments. The coefficient is >0.50, this suggests that there is a **moderately high correlation**.
- The second place goes to CharacTER.
- The third place goes to Levenshtein Edit Distance.
- TER shows a worse correlation compared to the 3 new metrics mentioned above.

# Italian

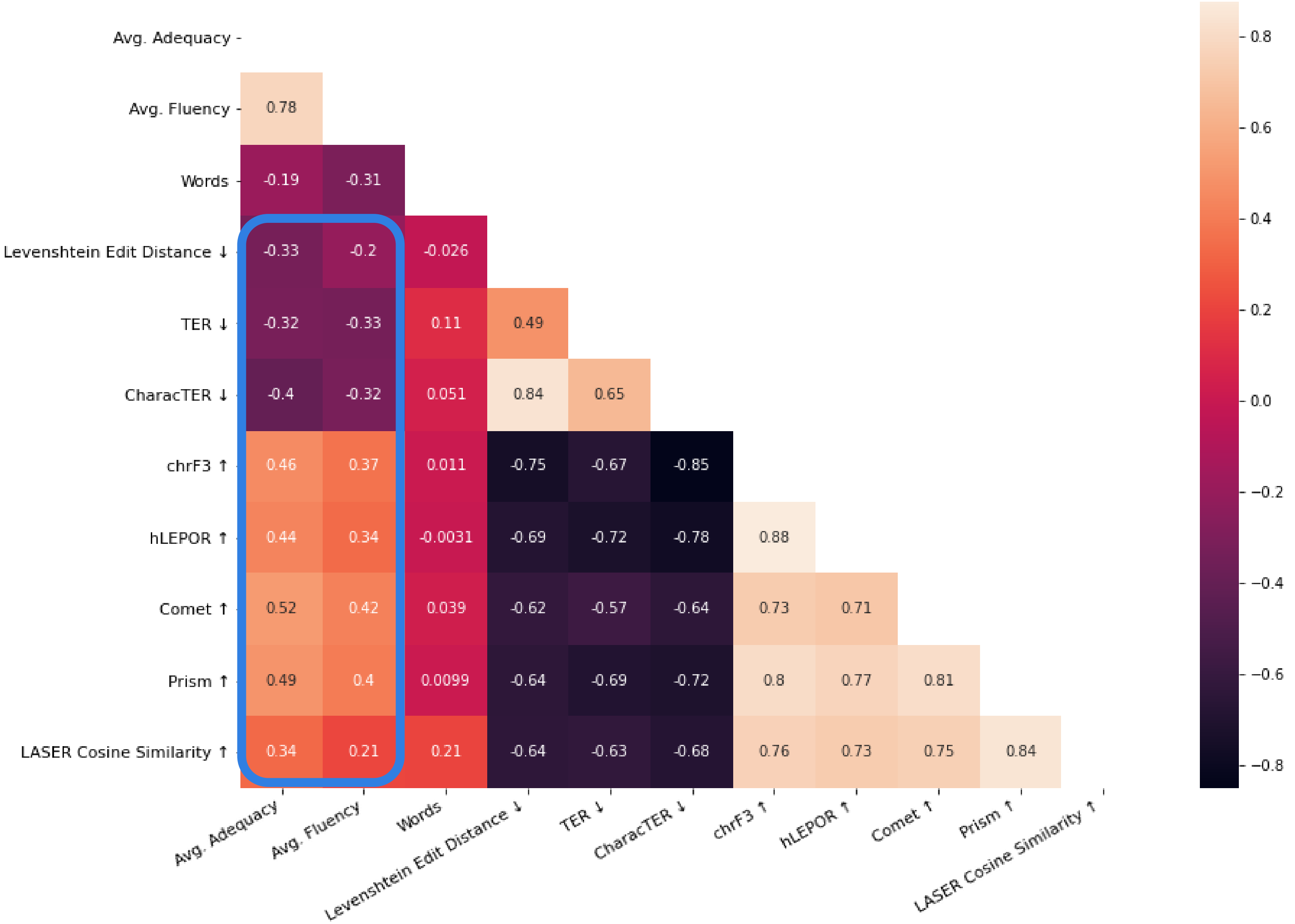



## Insights

Pearson Correlation Coefficient calculated to analyze the correlation between Human Assessment and metrics, as well as between each one of the metrics included in the study.

- The best correlation between Human Assessments and metric is seen with **COMET**.
- The second place goes to chrF3 and Prism, which show comparable results (chrF3 better correlates with Fluency, compared to Prism).
- The third place goes to CharacTER and hLEPOR, which show comparable results.
- Levenshtein Edit Distance and TER show a worse correlation compared to the 3 new metrics mentioned above.

# Russian

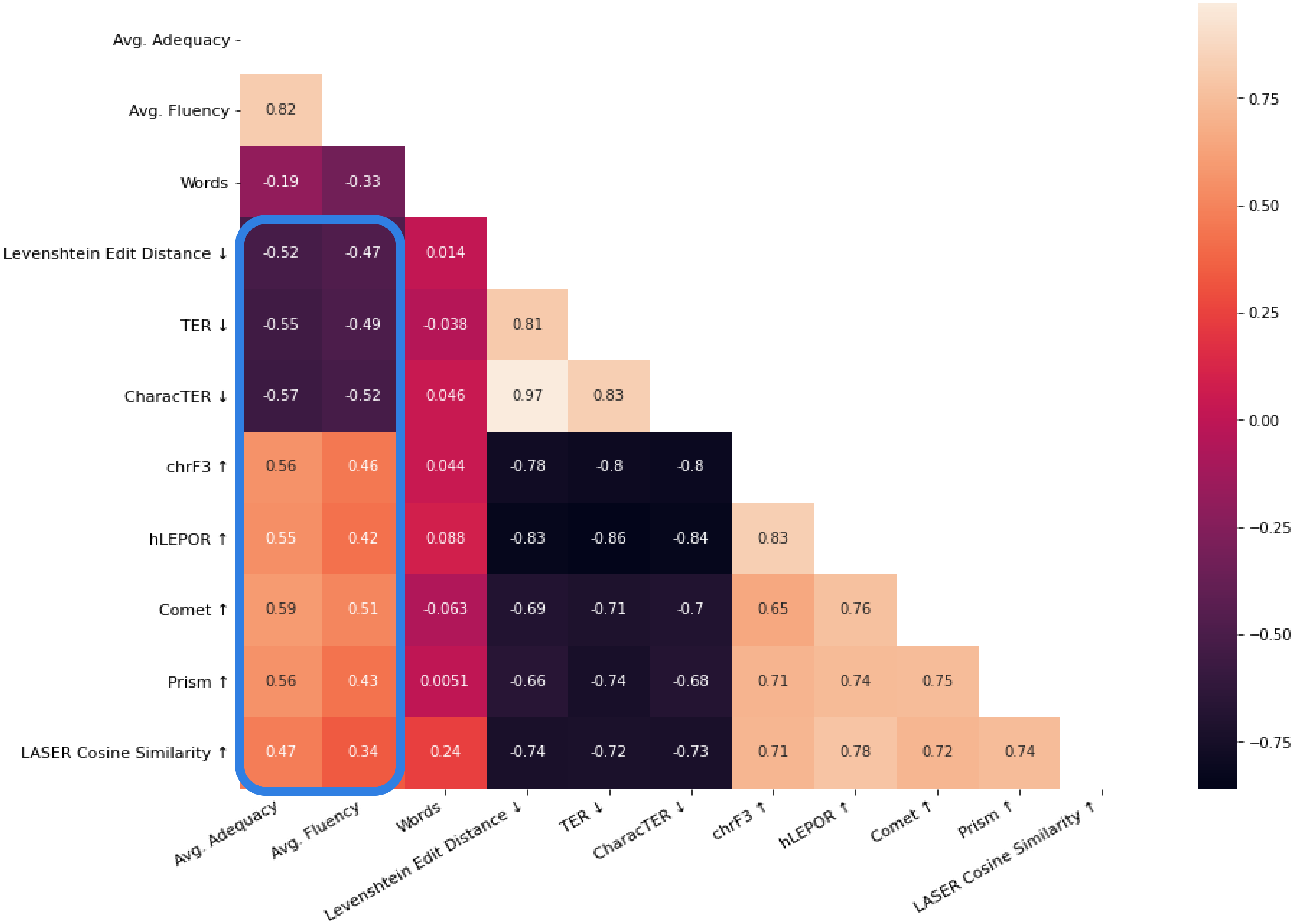



## Insights

Pearson Correlation Coefficient calculated to analyze the correlation between Human Assessment and metrics, as well as between each one of the metrics included in the study.

- **COMET** is the metric that achieves the best correlation with Human Assessments. The coefficient is >0.50 with Accuracy, this suggests that there is a **moderately high correlation**.
- The second place goes to Prism, which also shows a high correlation, close to 0.50.
- The third place goes to chrF3 and hLEPOR which show comparable results.
- Levenshtein Edit Distance and TER show a significantly worse correlation compared to the 3 new metrics mentioned above.

# Simplified Chinese




### Insights

Pearson Correlation Coefficient calculated to analyze the correlation between Human Assessment and metrics, as well as between each one of the metrics included in the study.

- **COMET** is the metric that achieves the best correlation with Human Assessments. The coefficient is >0.50, this suggests that there is a **moderately high correlation**.
- The second place goes to CharacTER, which show comparable results.
- The third place goes to Prism and hLEPOR, which also show a high correlation with Accuracy.
- Levenshtein Edit Distance and TER also show a good correlation.
- **Need to investigate why correlations are overall better for Chinese.**

# Conclusions

- Overall, **COMET achieves the highest correlation** with Human Assessment for each language (for some languages >0.50 Pearson correlation coefficient).
- **Prism, characTER and chrF3 also show good correlation** with Human Assessment across the board.
- **Laser Cosine Similarity score** is the only metric which shows a positive **correlation** (>0.20) **with the number of words** in the source segment for every language. This could suggest that Laser Cosine Similarity might does not perform well on shorter segments.
- **No significant differences were noticed in correlations based on the content type** (Product UI/UA vs Marketing). All metrics achieve at least moderate correlations ($\pm$ 0.30).
- **All the new metrics analyzed show a better correlation with Human Assessment** per language **compared to TER and Levenshtein Edit Distance**. Slightly different observation for Hindi.
- **Business implications**: ideally, the metric(s) with higher correlation should be used to evaluate the quality of the raw machine translation output, analyze the post-editing effort (which is closely related to MTPE discounts) and in quality estimation. Because we have seen that the preferred metric varies depending on the language, this could mean to have different "go-to" metrics in place, depending on the language in scope.

# Further Research

1. Test the metrics on more languages – what is the best metric for every language and why? Is it possible and convenient for an LSP to use different preferred metrics for every language?
2. Establish the acceptability threshold for the most relevant metrics – what is a good score and what is a bad score?
3. Get a better understanding of the reasons underlying variance of the same metric across different languages.

# Thank you

# And Special Thanks to…

Alex Yanishevsky

Anna Pizzolato

David Clarke

Elaine O’Curran

Jon Cambra

Lena Marg

welocalize

# Appendix

welocalize

# Metrics Definition

**Levenshtein Edit Distance**: The number of insertions, deletions, substitutions required to transform MT output to the human reference translation based on the Levenshtein algorithm. In our analysis, we normalize this value by the number of characters in the MT output.

**TER** (Translation Edit Rate): is a word-based error metric for machine translation that measures the number of edits (insertions, deletions, substitutions and shifts) required to change a system output into one of the human references.

$$\mathrm{TER} = \frac{\text{\# of edits}}{\text{average \# of reference words}}$$

**CharacTER**: same as TER, but insertions, deletions, substitutions are calculated at the character level. The shift edit operation is still performed at word level. Unlike TER, the edit distance is normalized by the length of the MT output.

$$\mathrm{CharacTER} = \frac{\text{shift cost} + \text{edit distance}}{\text{\#characters in the hypothesis sentence}} \quad (1)$$

**chrF3**: F3 score based on character n-grams of size 6. The F3 score can be defined as the harmonic mean of precision and recall, with recall having three times more weight than precision (β = 3)

$$\mathrm{CHRF}\beta = (1+\beta^2)\frac{\mathrm{CHRP}\cdot\mathrm{CHRR}}{\beta^2\cdot\mathrm{CHRP}+\mathrm{CHRR}} \quad (1)$$

# Metrics Definition

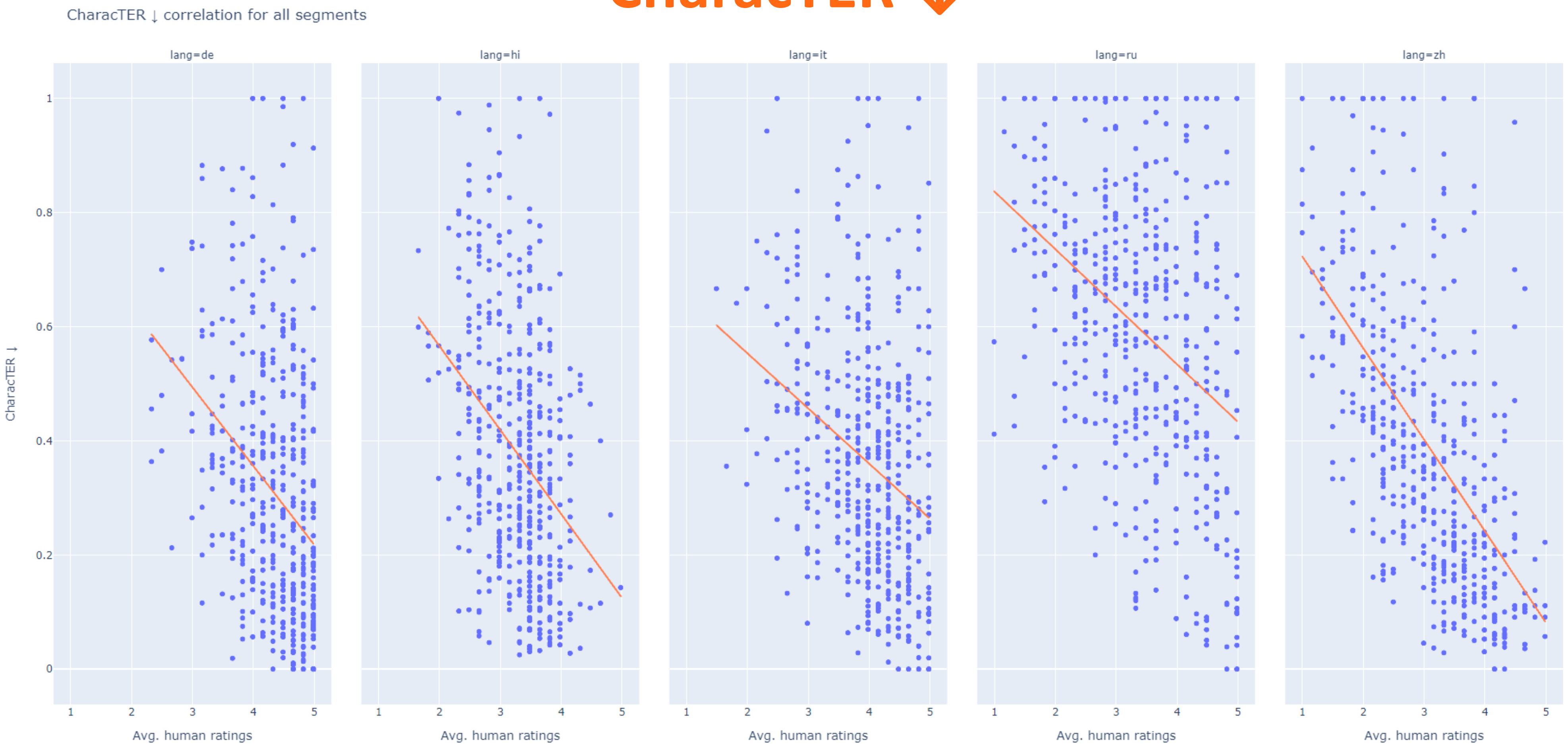
**hLEPOR**: computes the similarity of n-grams between a MT output and a reference translation, taking into account a length penalty, an n-gram position difference penalty, and recall.

**COMET**: a framework to train multilingual MT evaluation models that can function as metrics. For our analysis, we used the publicly available wmt-large-da-estimator-1719 model, which is trained to predict human judgments from WMT by leveraging sentence embeddings extracted from the source, MT output and reference segment.

**Prism**: uses a multilingual NMT system to score MT outputs conditioned on their corresponding human references. The score is calculated by averaging the log-probability for each token in the output assigned by the model.

**LASER cosine similarity**: LASER is a neural model trained on parallel data from 93 languages open sourced by Facebook in 2019. Sentence embeddings produced by its encoder can be compared to measure intra or interlingual semantic similarity using cosine similarity.

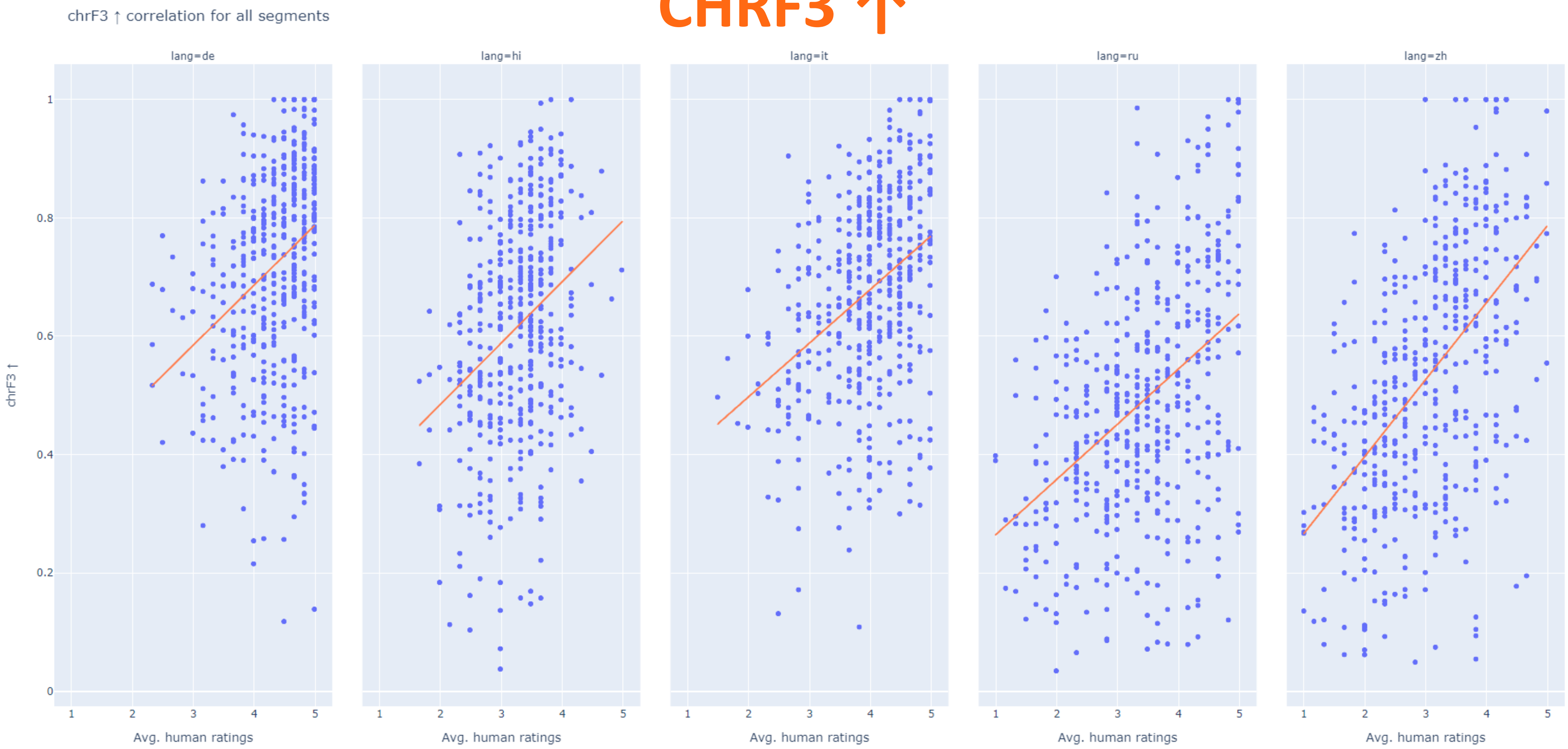
# CharacTER ↓




**Key:**

- Avg. Human ratings = Adequacy and Fluency ratings by 3 linguists averaged per segment
- Trendline = the degree to which Avg. Human ratings and CharacTER scores are correlated. A diagonal line indicates a perfect correlation. The more points close to the line, the stronger the correlation.

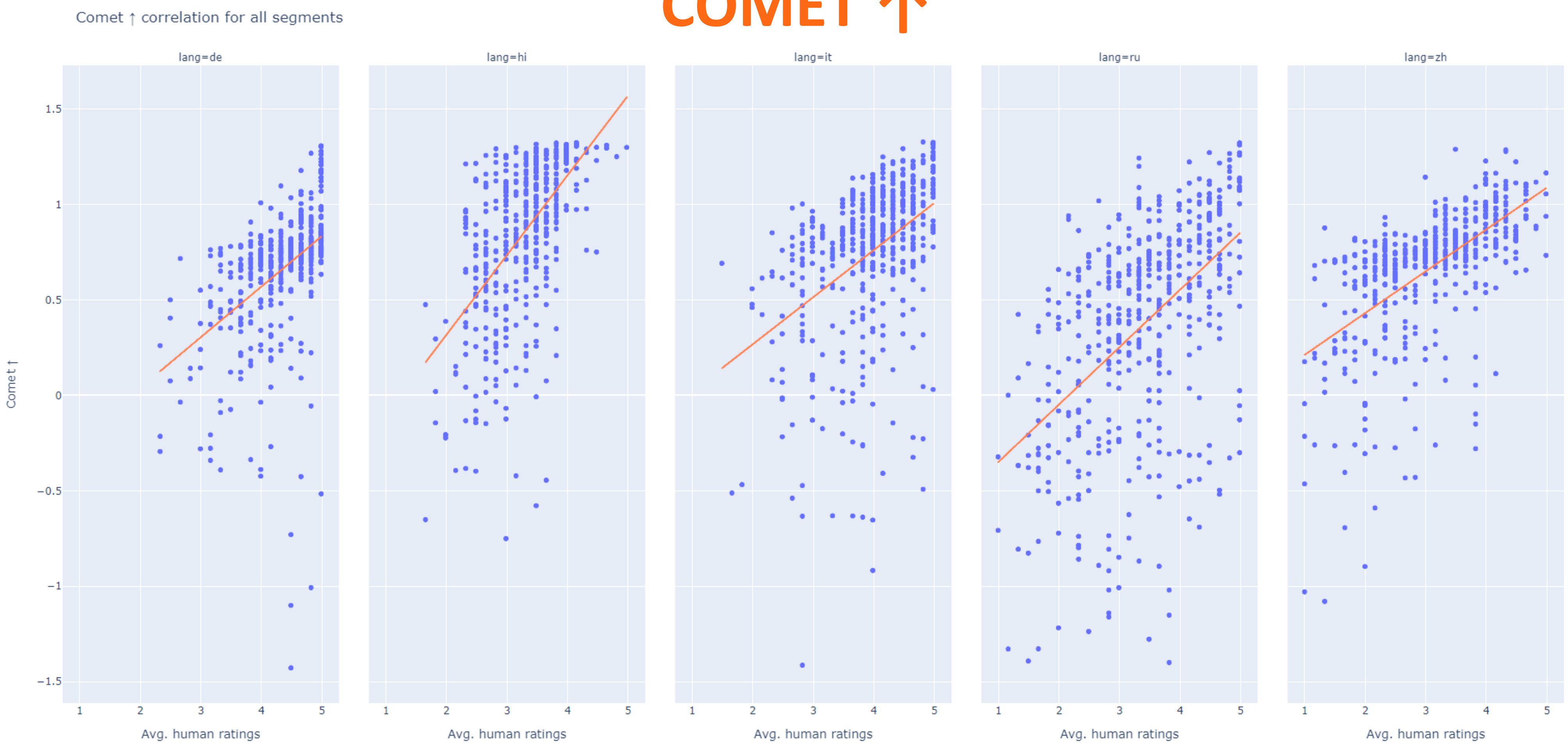
# CHRF3 ↑




**Key:**

- Avg. Human ratings = Adequacy and Fluency ratings by 3 linguists averaged per segment
- Trendline = the degree to which Avg. Human ratings and chrF3 scores are correlated. A diagonal line indicates a perfect correlation. The more points close to the line, the stronger the correlation.

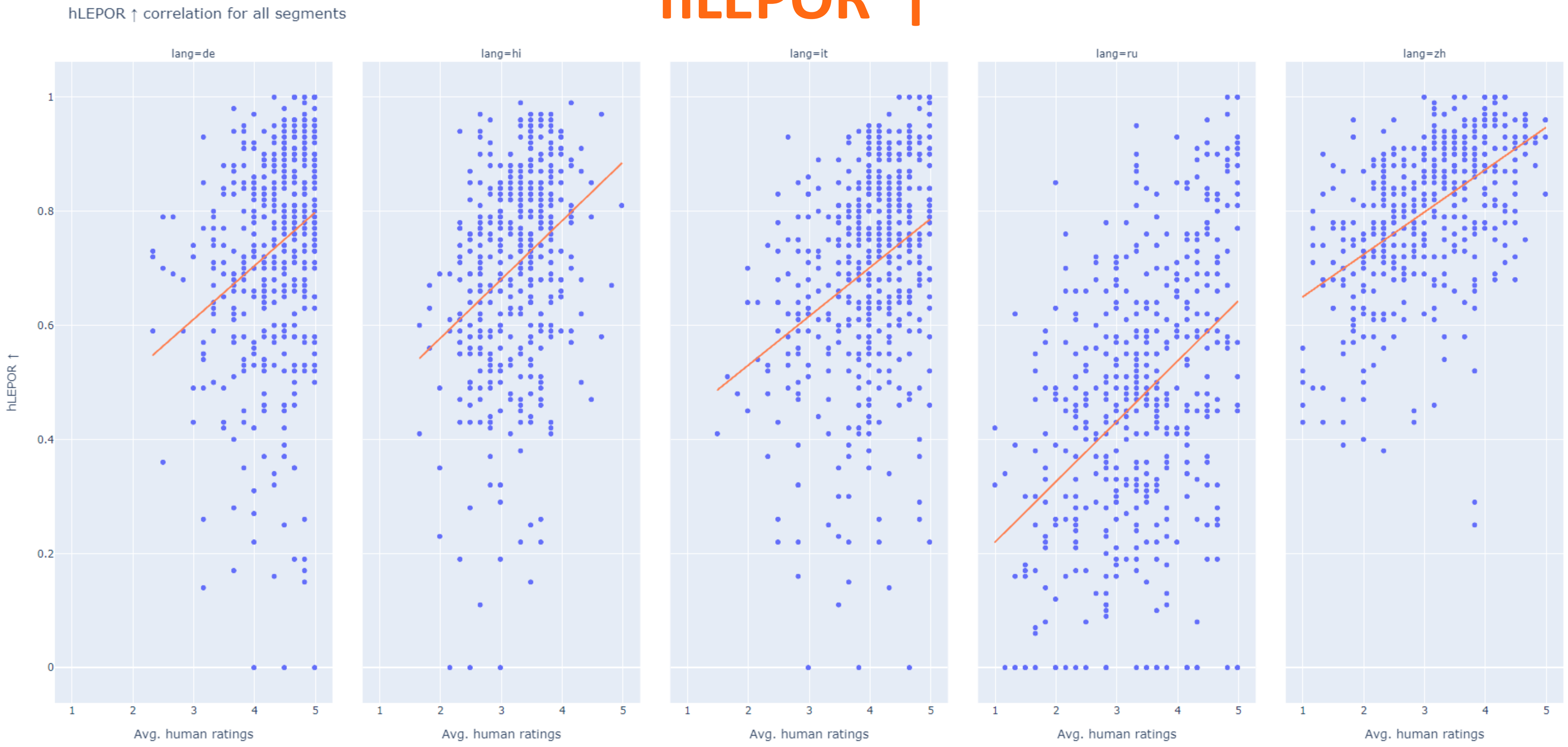
# COMET ↑




**Key:**

- Avg. Human ratings = Adequacy and Fluency ratings by 3 linguists averaged per segment
- Trendline = the degree to which Avg. Human ratings and COMET scores are correlated. A diagonal line indicates a perfect correlation. The more points close to the line, the stronger the correlation.

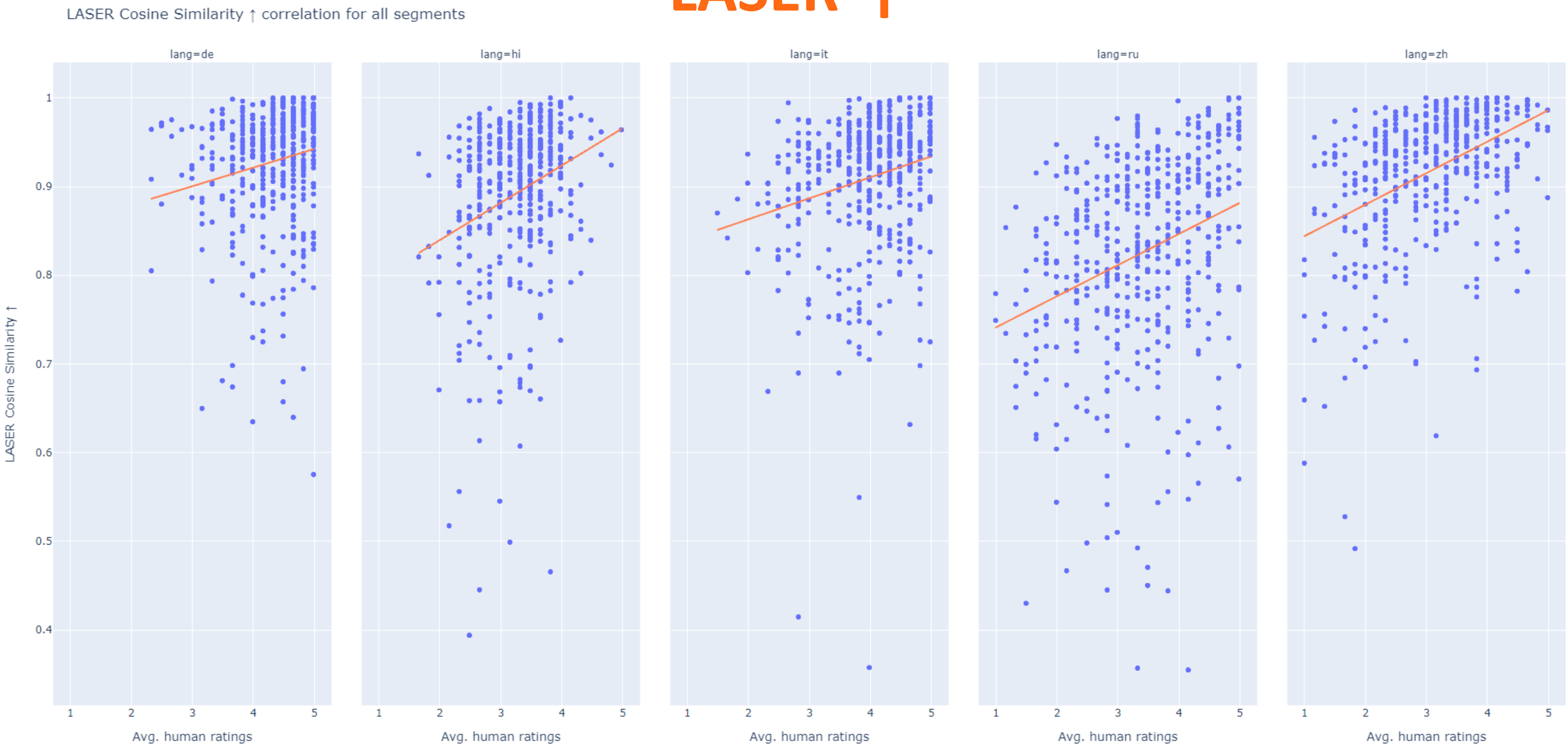
# hLEPOR ↑




**Key:**

- Avg. Human ratings = Adequacy and Fluency ratings by 3 linguists averaged per segment
- Trendline = the degree to which Avg. Human ratings and hLEPOR scores are correlated. A diagonal line indicates a perfect correlation. The more points close to the line, the stronger the correlation.

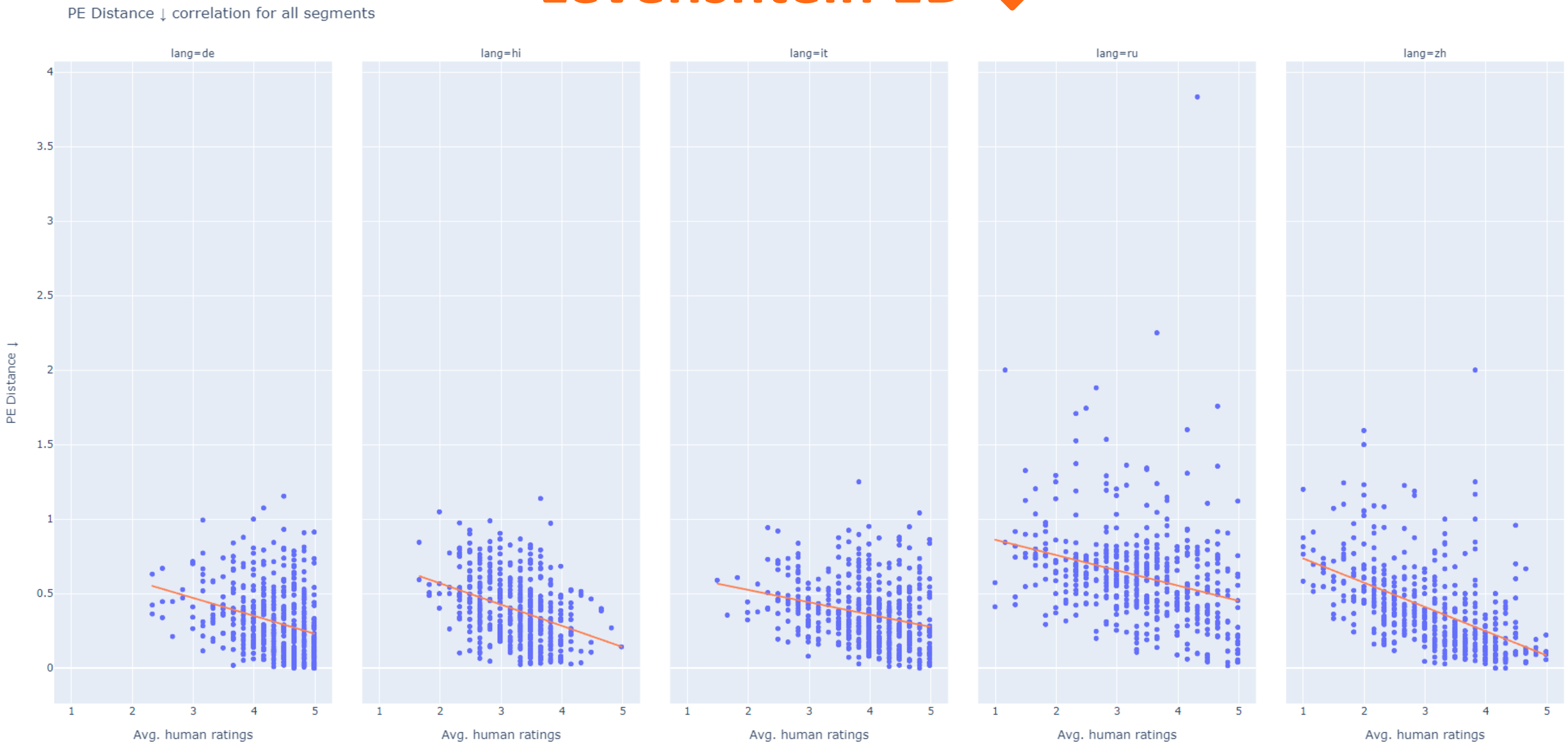
# LASER ↑




**Key:**

- Avg. Human ratings = Adequacy and Fluency ratings by 3 linguists averaged per segment
- Trendline = the degree to which Avg. Human ratings and LASER cosine similarity scores are correlated. A diagonal line indicates a perfect correlation. The more points close to the line, the stronger the correlation.

# Levenshtein ED ↓




**Key:**

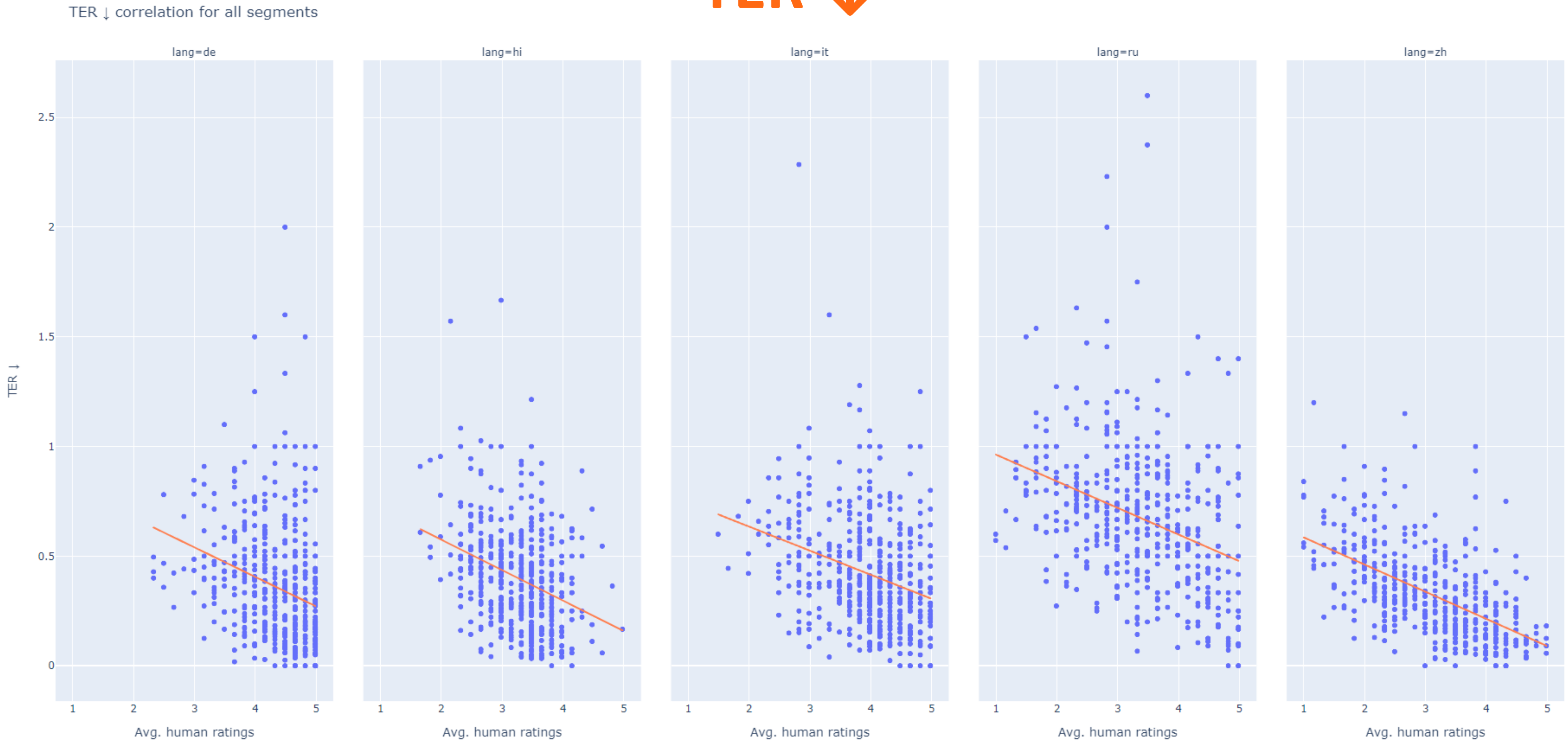
- Avg. Human ratings = Adequacy and Fluency ratings by 3 linguists averaged per segment
- Trendline = the degree to which Avg. Human ratings and Levenshtein Edit Distance scores are correlated. A diagonal line indicates a perfect correlation. The more points close to the line, the stronger the correlation.

# PRISM ↑


**Key:**

- Avg. Human ratings = Adequacy and Fluency ratings by 3 linguists averaged per segment
- Trendline = the degree to which Avg. Human ratings and PRISM scores are correlated. A diagonal line indicates a perfect correlation. The more points close to the line, the stronger the correlation.

# TER ↓




**Key:**

- Avg. Human ratings = Adequacy and Fluency ratings by 3 linguists averaged per segment
- Trendline = the degree to which Avg. Human ratings and TER scores are correlated. A diagonal line indicates a perfect correlation. The more points close to the line, the stronger the correlation.

# References

Artetxe, Mikel and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *arXiv:1812.10464 [cs]*

Banerjee, S. and Lavie, A., 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65–72.

Coughlin, D., 2001. Correlating Automated and Human Assessments of Machine Translation Quality.

Doddington, G., 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *HLT '02: Proceedings of the second international conference on Human Language Technology Research*, 138–145.

Han, Lifeng, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013. Language-independent Model for Machine Translation Evaluation with Reinforced Factors.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395.

*Proceedings of the 5th Conference on Machine Translation (WMT)*, pages 1–55, November 19–20, 2020.

Rei, Ricardo, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers,* 223–231.

# References

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers,* 223–231.

Thompson, Brian and Matt Post. 2020. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. *arXiv:2004.14564 [cs]*

Wang, Weiyue, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation Edit Rate on Character Level. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. 505–510.