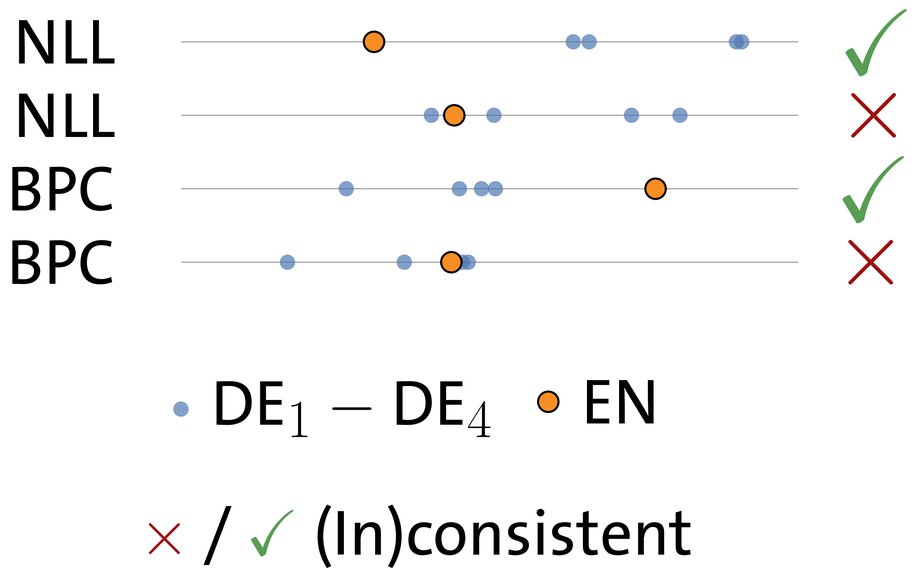
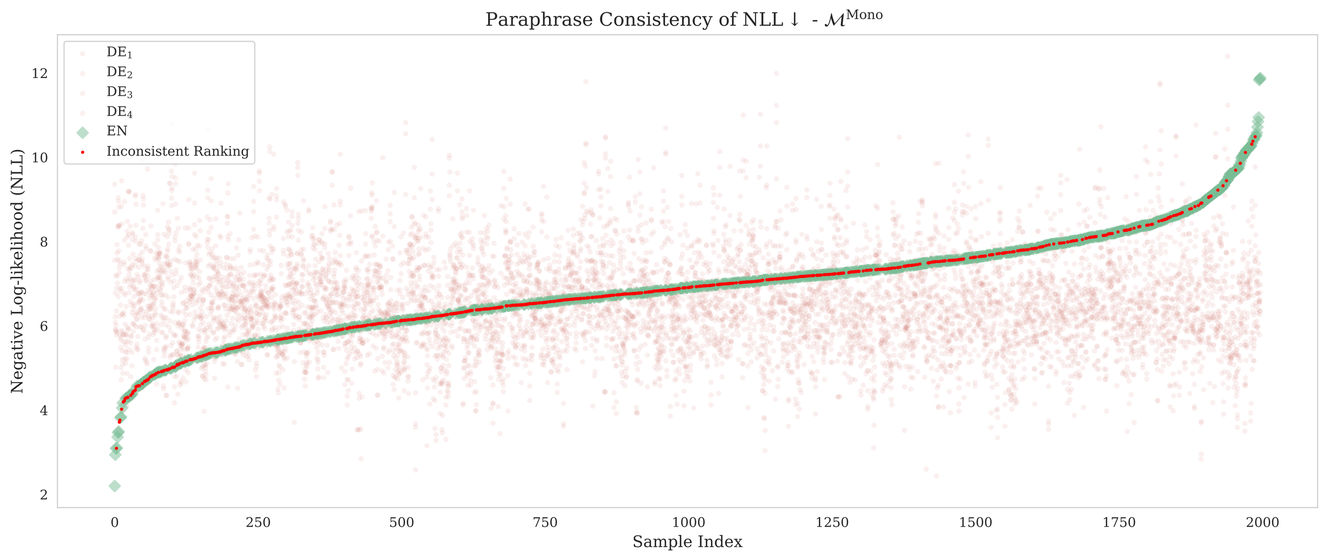
Form and Meaning in Intrinsic Multilingual Evaluations

Wessel Poelman & Miryam de Lhoneux
LAGoM-NLP, Department of Computer Science, KU Leuven
wessel.poelman@kuleuven.be





1. Background

- "Which language is hard to language model?" has received quite some attention.
- Intrinsic metrics like perplexity are often used to answer such questions.
- Evaluations use **multi-parallel** datasets; presumed to allow for a fair comparison.
  - Examples: FLORES-200, EuroParl, UN Parallel Corpus, Parallel Bible Corpus, ...
- Are metrics that operate on **form** fair when they have consistent **meaning**?

2. Consistency

Consistency of language models:
when presented with paraphrases, is the output still the same?

Model	Sequence	PPL
A	Sabe jugar al ajedrez	20
B	Do you know how to play chess	22
B	Can you play chess	18

- Sentences have (roughly) the same *meaning*.
- PPL measures *form*. Results in two issues:
  1. Within a language, values can be different.
  2. Across languages, comparisons are affected.
- Conclusions can flip based on which test set we pick!

3. Setup

- Two setups: multilingual models & multiple monolingual models.
- Pre-training data: EuroParl and UN Parallel Corpus, **both multi-parallel**.
- Evaluation: FLORES-200 and WMT-19 paraphrase (parallel, ∼2000 samples).
- WMT-19: one English sentence with **four** human translated German sentences.

4. Metrics

All used in previous works; various reasons of why comparisons should be fair. For a sequence S, segmented into w (tokens), c (characters):

$$\text{Negative Log-Likelihood (NLL)} = -\frac{1}{S}\sum_{t=1}^{S} \log P(w_t|w_{<t})$$

$$\text{Perplexity (PPL)} = \exp(\text{NLL})$$

$$\text{Bits per Character (BPC)} = -\frac{1}{S}\sum_{t=1}^{S} \log_2 P(c_t|c_{<t})$$

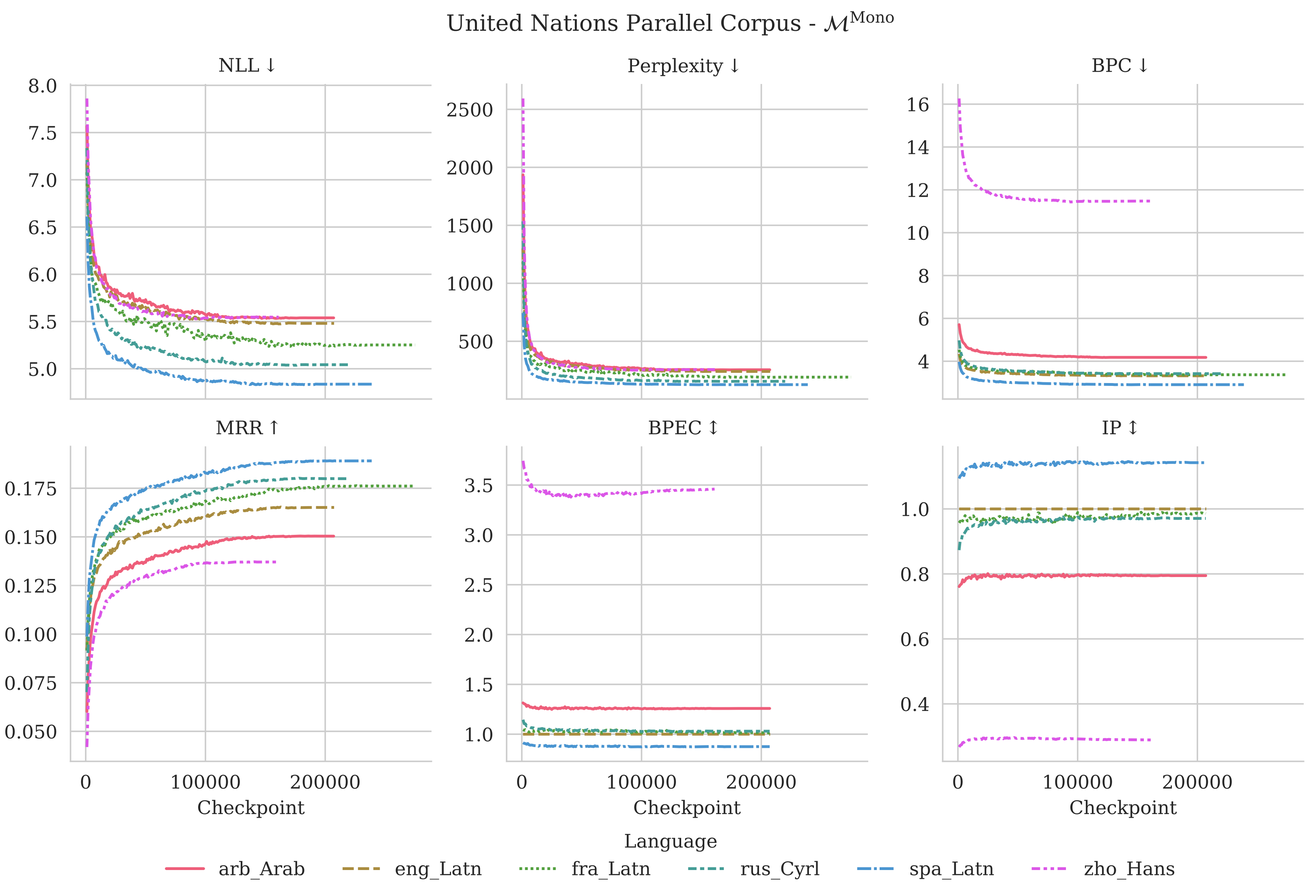
$$\text{Bits per English Character (BPEC)} = \frac{\text{BPC}_{\text{Target}}}{\text{BPC}_{\text{EN}}}$$

$$\text{Information Parity (IP)} = \frac{\text{BPC}_{\text{EN}}}{\text{BPC}_{\text{Target}}}$$

$$\text{Mean Reciprocal Rank (MRR)} = \frac{1}{S}\sum_{t=1}^{S} \frac{1}{R_t}$$ where R is the ranking over the vocabulary

5. Results

5.1 Tempting to Draw Conclusions...



5.2 But What About the Consistency?

Metric	$\mathcal{M}^{\text{Mono}}$	$\mathcal{M}^{\text{Multi}}$
NLL	47%	51%
BPC	50%	52%
MRR	50%	51%

Sample-level inconsistencies.

Model	Metric	DE^{F}	EN	DE^{S}
$\mathcal{M}^{\text{Mono}}$	NLL ↓	6.43–6.64 ✓	6.91	5.89–7.23 ×
	BPC ↓	2.09–2.20 ✓	2.27	1.82–2.44 ×
	MRR ↑	19.6–21.2 ✓	18.6	14.6–26.0 ×
$\mathcal{M}^{\text{Multi}}$	NLL ↓	6.81–7.02 ✓	7.05	6.25–7.60 ×
	BPC ↓	2.40–2.51 ✓	2.54	2.13–2.77 ×
	MRR ↑	22.5–23.7 ✓	21.4	17.4–28.7 ×

Split-level inconsistency, with row-wise sorting.

6. Conclusion

- Consistency in *meaning* does not neutralize differences in **form**.
- Metrics that transform the NLL are inherently measuring (and sensitive to) form.
- **Within** languages: not the same values for paraphrases, **across** languages: many inconsistent samples.

Contact & Acknowledgments

WP is funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096.