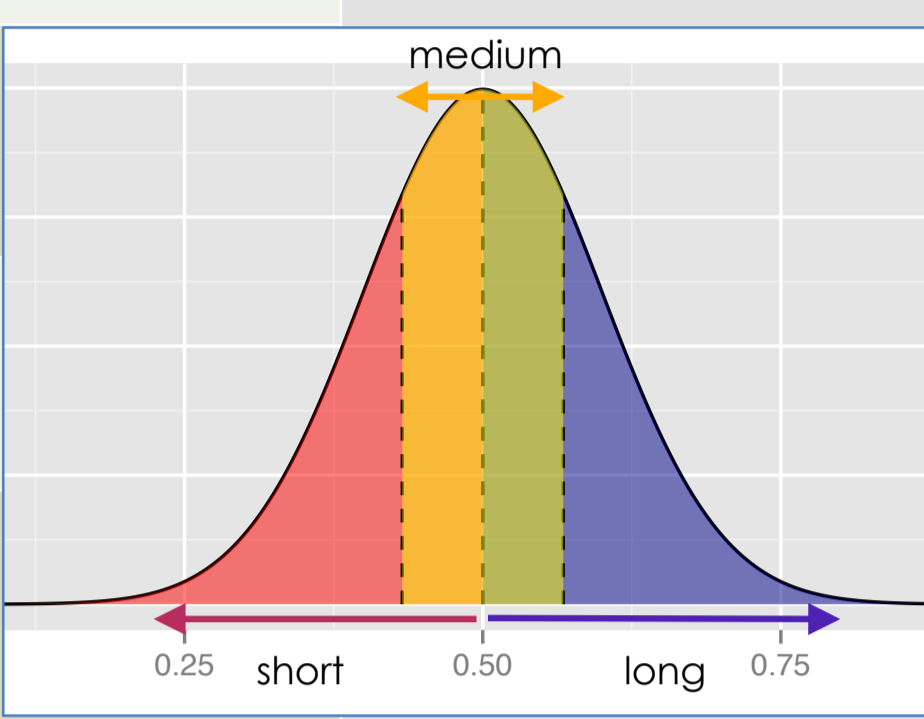


1. Introduction

- MT performance depends on the **tuning set** (Zheng et al., 2010, Liang et al 2010)
- Optimization can be improved by selecting a suitable tuning set.
- PRO has *issues* with **length**:
 - generates shorter translations (Nakov et al., 2012)
 - is susceptible to produce pathologically long translations (Nakov et al., 2013)

3. Setup

	Arabic-English	Spanish-English
Datasets	NIST 04, 05, 06, 09	WMT 08, 09, 10, 11
References	multiple, single	single
Partitions	Length: short, mid, long Verbosity: low, mid, high	
Optimizers	PRO & MERT	

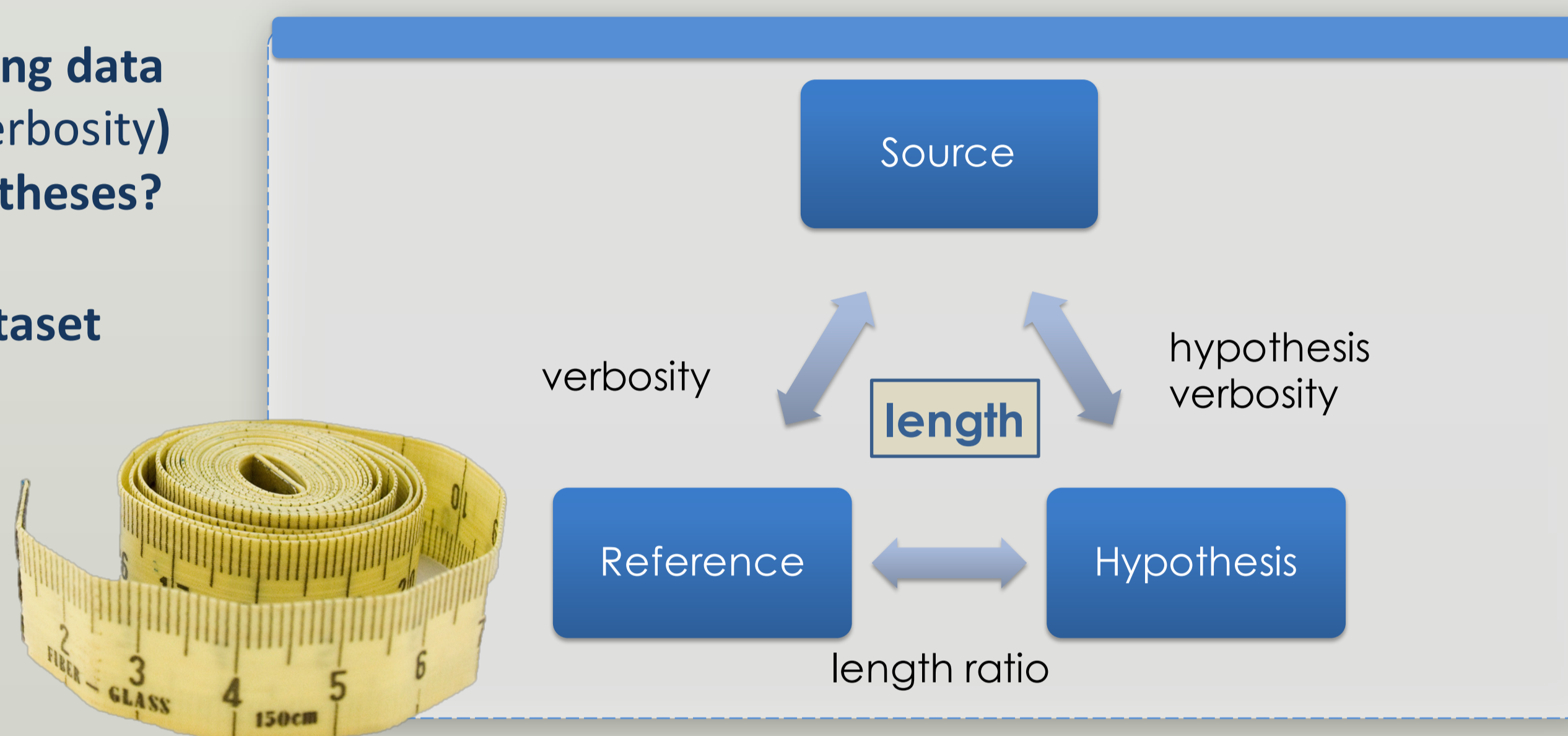


PRO
vs
MERT

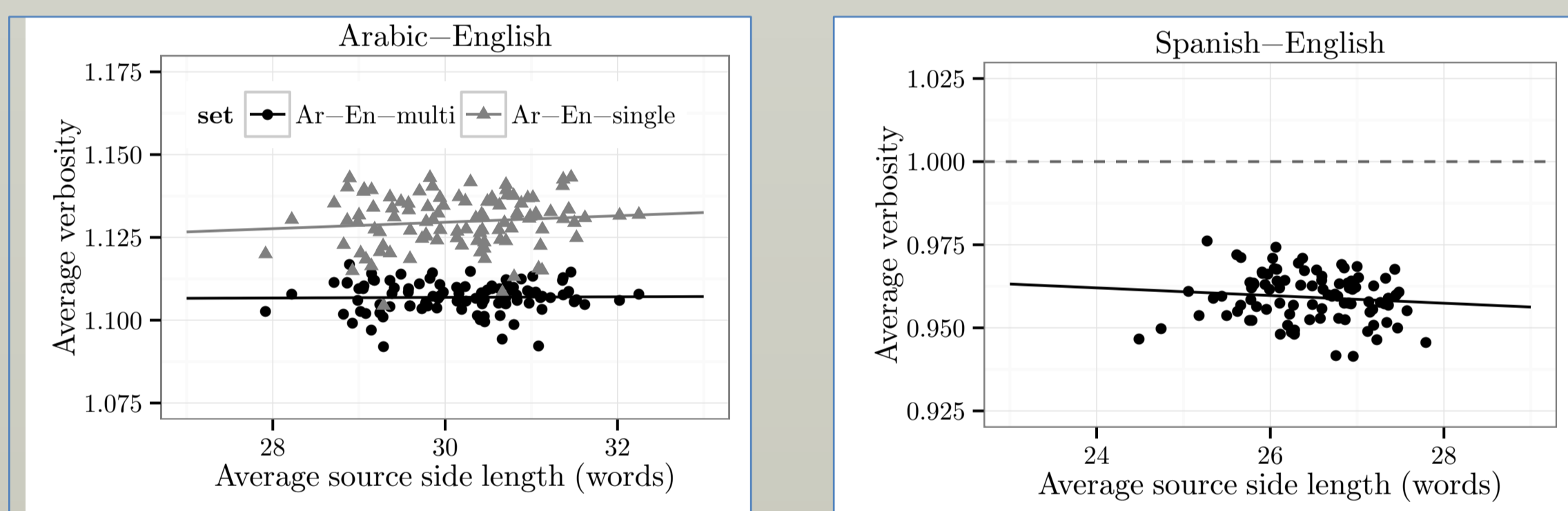
2. Analyzing Length for SMT Optimization

The effect of tuning data (source length, verbosity) on the *final hypotheses*?

What type of dataset works best?



Tuning set verbosity depends on source length



Arabic: Target (English) sentences have more words than source sentences (verb>1), and they get longer with longer sentences.

Spanish: Target sentences have fewer words. They get shorter with longer source sentences.

What we discovered

Source length <-> verbosity
We can use source length to control verbosity

4. Results: Length

MERT

98%

Correlation: Tuning verbosity vs. hypothesis verbosity

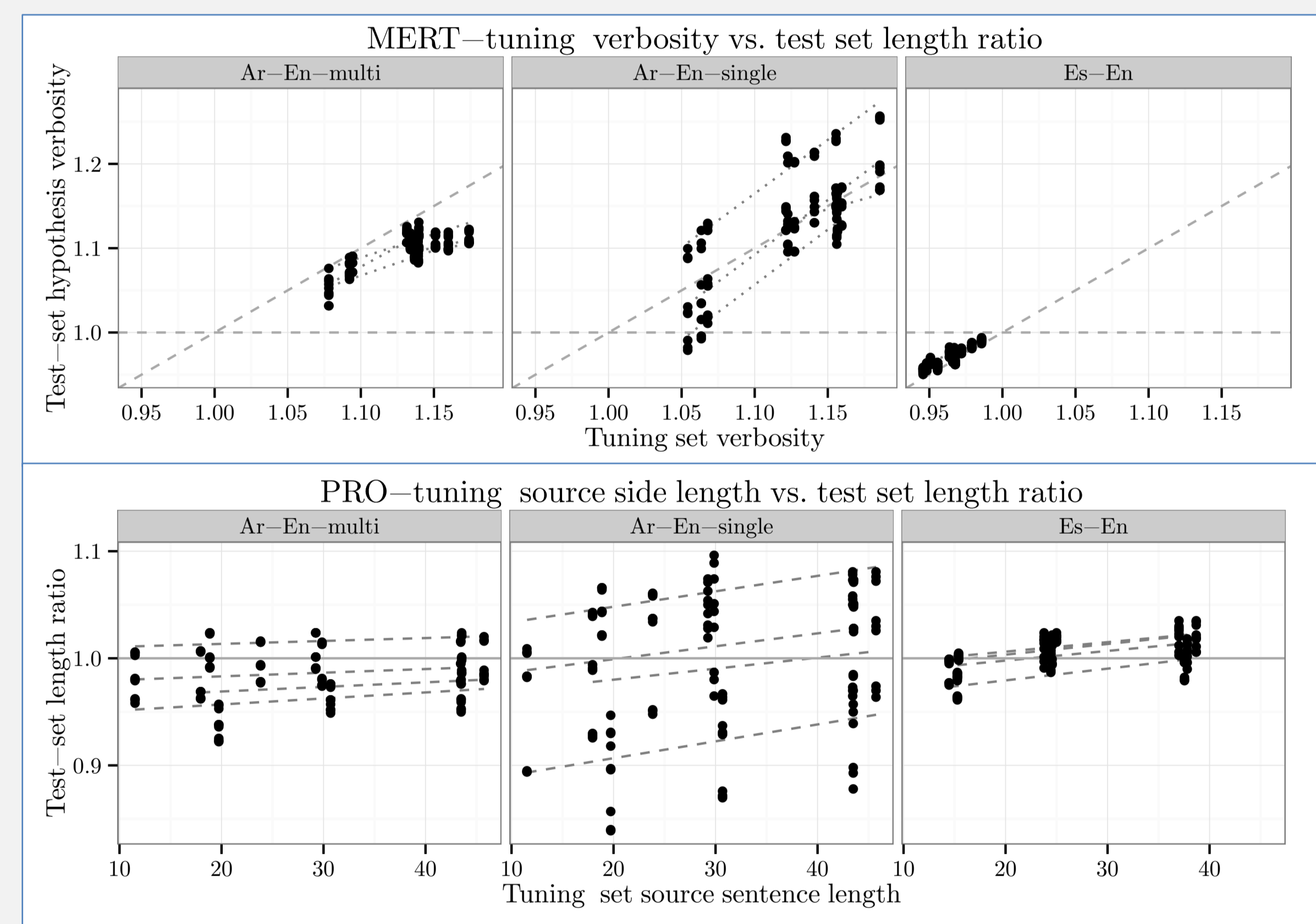
PRO

44%

Correlation: Tuning verbosity vs. hypothesis verbosity

67%

Correlation: Tuning source length vs. length ratio

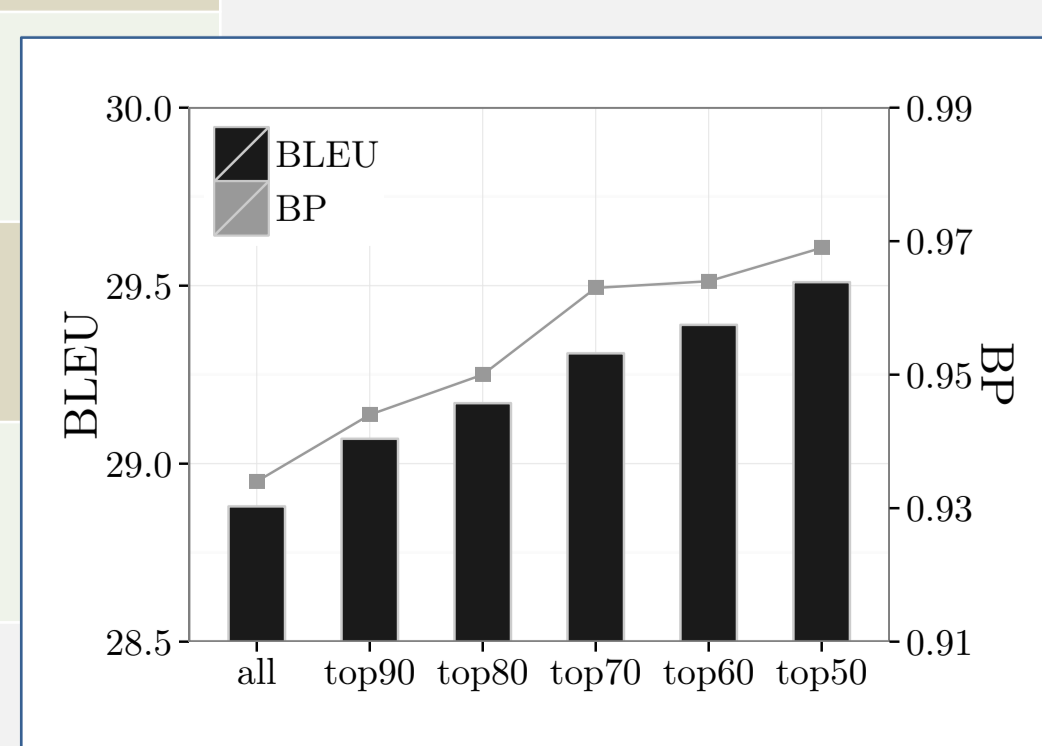


6. Choosing a Tuning Set: verbosity or length?

tuning	test					
	Arabic-English (multi-ref)		Arabic-English (1-ref)		WMT Spanish-English	
	MERT	PRO-fix	MERT	PRO-fix	MERT	PRO-fix
length						
short	48.71	49.12	26.74	27.35	26.79	27.07
mid	49.27	49.59	26.97	27.23	26.99	26.88
long	49.35	49.20	27.23	27.28	27.02	26.84
verbosity						
low-verb	47.90	47.60	25.89	25.88	26.70	26.61
mid-verb	49.16	49.52	27.69	27.95	27.09	26.81
high-verb	50.28	50.79*	27.36	28.03*	27.01	27.38*

- **MERT:** Choose long source tuning sets!
- **PRO:** Choose high-verbosity tuning sets!
- Avoid low-verbosity sets
- Prefer verbosity as a selection criteria

	PRO	MERT
Likes	Length	Verbosity
Best strategy	High verbosity tuning set	Mixed. High verbosity tuning set
Worst strategy	Select low verbosity tuning sets	



5. Results: BLEU

tuning	Arabic-English (1-ref)		
	short	mid	long
MERT			
short	26.69*	28.14	27.49
mid	26.22	28.39*	27.96
long	25.80	28.20	28.27*
PRO-fix			
short	25.95	27.66	27.28
mid	25.98	28.23	28.19
long	25.87	28.11	28.05

Cross-testing (BLEU)

MERT: best BLEU when tuning on similar-to-test

PRO: learned parameters are independent of test-set length.

7. Conclusion

- **Know your tuning datasets:** Different language pairs and translation directions may have different *source-side length* – *verbosity* dependencies.
- **When optimizing with PRO:** select or construct a high-verbosity dataset as this could potentially compensate for PRO's tendency to yield too short translations.
- **When optimizing with MERT:** If you know beforehand the test set, select the *closest* tuning set. Otherwise, tune on longer sentences.