Test Set Influence on BLEU

Havlíček (2007) evaluates the influence of:

- number of reference translations,
- translation direction.

on *human-produced* text (1 human translation against 4 others).

	cs→en, Professionals				en→cs, Math Students			
Refs	Indiv. Results			**Avg**	Indiv. Results			**Avg**
1	41.15	32.66	34.03	**35.95**	3.66	8.62	5.79	**6.02**
2	49.09	49.78	41.26	**46.71**	9.82	8.26	9.36	**9.15**
3	52.63			**52.63**	13.06			**13.06**

⇒ heavy dependence on the number of references.

More references allow to match more n-grams of MT output.

⇒ heavy dependence on the translation direction and quality.