Contradictions in Manual Evaluation

Results for WMT10:

Evaluation Method	Google	CU-Bojar	PC Translator	TectoMT
$\geq$ others (WMT10 official)	**70.4**	65.6	62.1	60.1
$>$ others	49.1	45.0	**49.4**	44.1
Edits deemed acceptable [%]	**55**	40	43	34
Quiz-based evaluation [%]	80.3	75.9	80.0	**81.5**
Automatic: BLEU	**0.16**	0.15	0.10	0.12
Automatic: NIST	**5.46**	5.30	4.44	5.10

Results for WMT19:

- Best systems match humans in GCSE-like scoring.
- They score worse in pseudo-doc-aware DA.
- They are absolutely terrible on agreements.

... each technique provides a different picture.