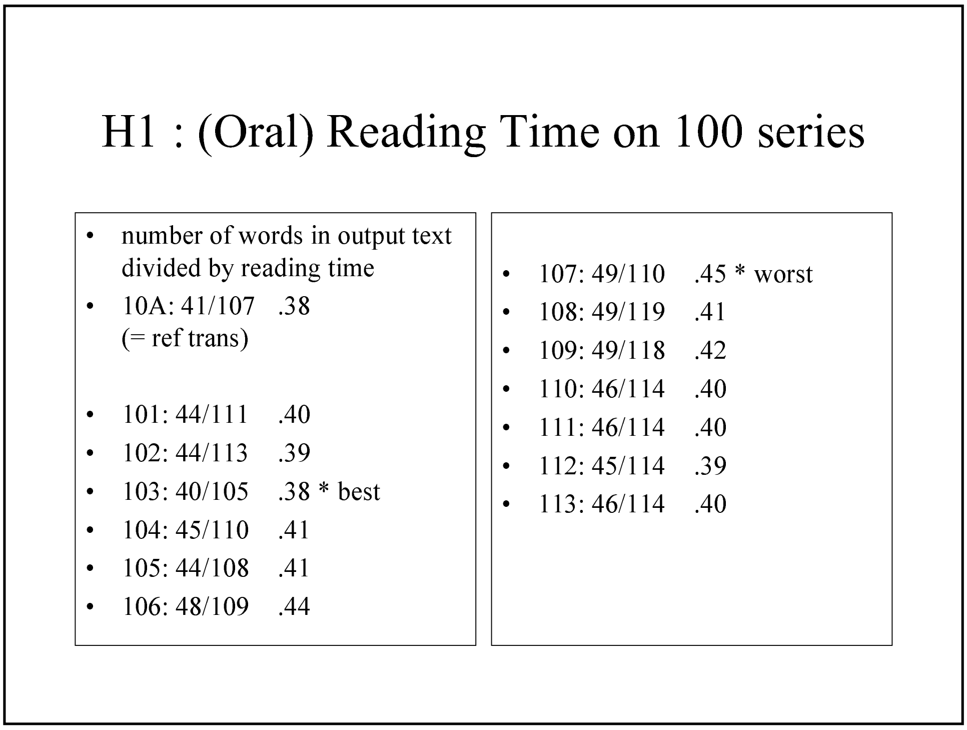
Report by Michelle Vanni

Prepared for the LREC 2002 workshop
on MT evaluation, 27 May 2002

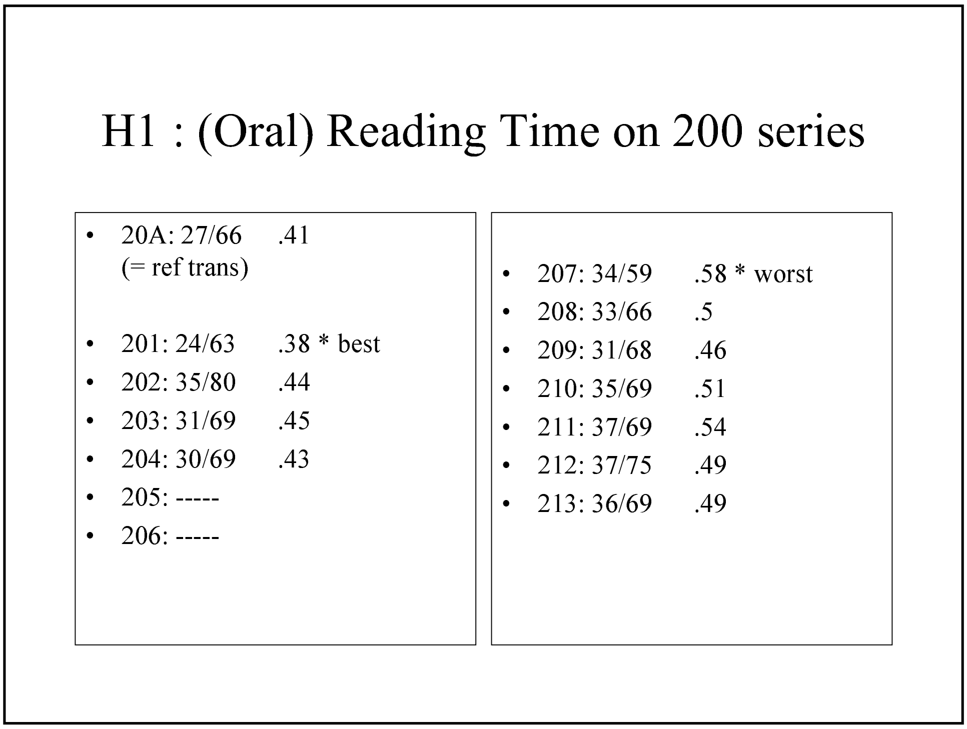
Metrics that were applied

- Two human-based measures (one evaluator)
- Performed on two sets of data
  - initial paragraphs of the outputs of the two source texts
  - 'Children and Drugs': 108 words of source text (<<Prévenir ... elle est efficace>>)
  - 'Taliban and Women':70 words of source text (<<Les décrets ... la seule capitale>>)



H1 : (Oral) Reading Time on 100 series

- number of words in output text divided by reading time
- 10A: 41/107 .38 (= ref trans)
- 101: 44/111 .40
- 102: 44/113 .39
- 103: 40/105 .38 * best
- 104: 45/110 .41
- 105: 44/108 .41
- 106: 48/109 .44
- 107: 49/110 .45 * worst
- 108: 49/119 .41
- 109: 49/118 .42
- 110: 46/114 .40
- 111: 46/114 .40
- 112: 45/114 .39
- 113: 46/114 .40



H1 : (Oral) Reading Time on 200 series

- 20A: 27/66 .41 (= ref trans)
- 201: 24/63 .38 * best
- 202: 35/80 .44
- 203: 31/69 .45
- 204: 30/69 .43
- 205: -----
- 206: -----
- 207: 34/59 .58 * worst
- 208: 33/66 .5
- 209: 31/68 .46
- 210: 35/69 .51
- 211: 37/69 .54
- 212: 37/75 .49
- 213: 36/69 .49

H2 : Correction (post-editing) time on 100 series

- (number of minutes spent in correction) / (total number of words in text) x 10

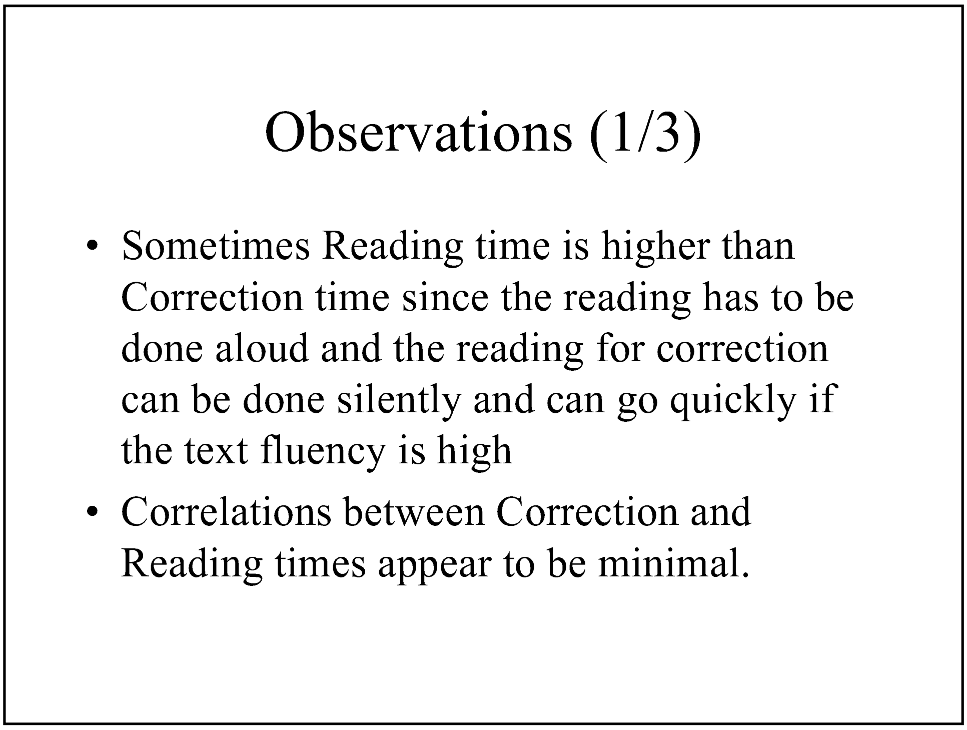
- 101: 90/44 20.50
- 102: 80/44 18.20
- 103: 75/40 18.75
- 104: 60/45 13.33
- 105: 32/44 7.38* best
- 106: 75/48 15.63

- 107: 155/49 31.63* worst
- 108: 155/49 31.63* worst
- 109: 86/49 17.55
- 110: 90/46 19.57
- 111: 75/46 16.30
- 112: 105/45 23.33
- 113: 75/46 16.30

H2 : Correction (post-editing) time on 200 series

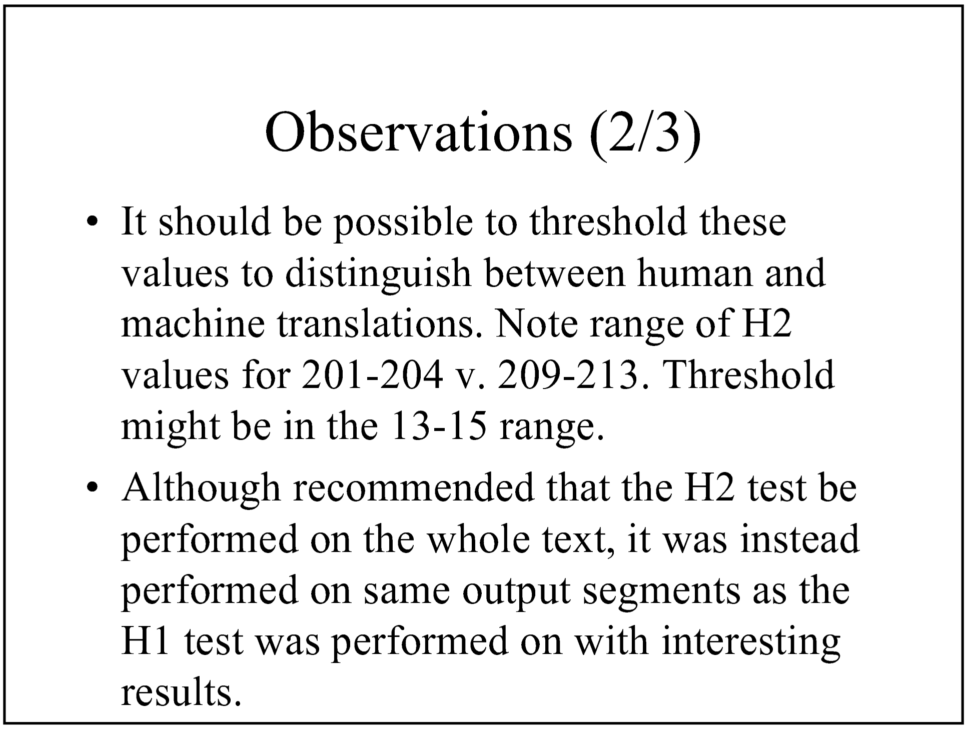
- 201: 25/24 10.42
- 202: 35/35 10. * best
- 203: 35/31 11.29
- 204: 30/30 10. * best
- 205: -----
- 206: -----

- 207: 180/34 52.94 * worst
- 208: 100/33 30.30
- 209: 105/31 33.87
- 210: 60/35 17.14
- 211: 105/37 28.38
- 212: 105/37 28.38
- 213: 80/36 22.22



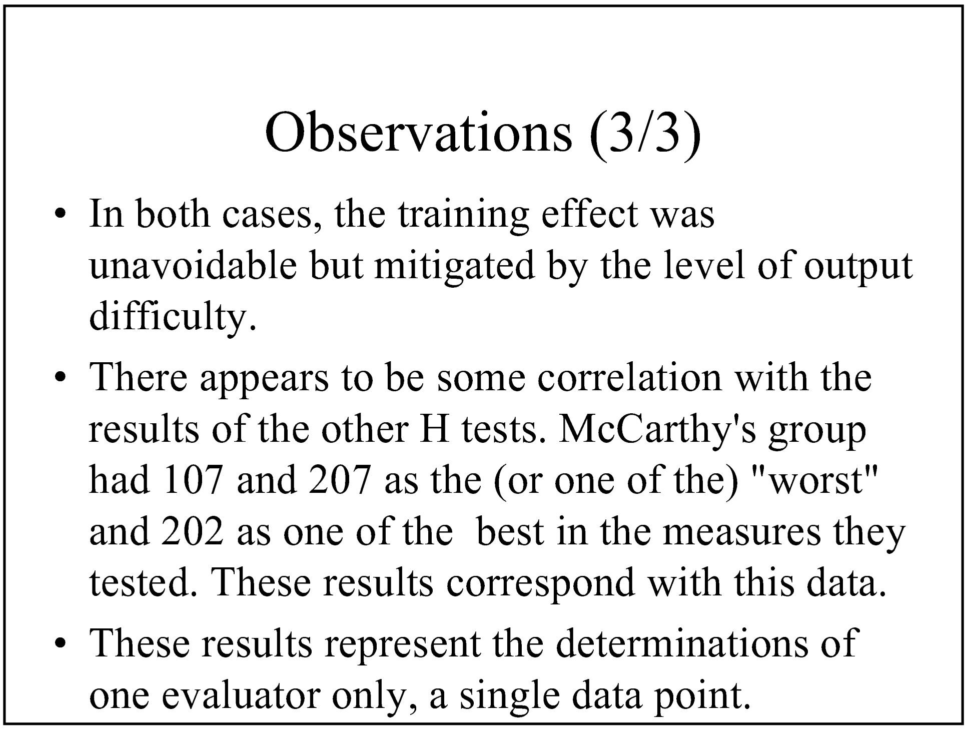
Observations (1/3)

- Sometimes Reading time is higher than Correction time since the reading has to be done aloud and the reading for correction can be done silently and can go quickly if the text fluency is high
- Correlations between Correction and Reading times appear to be minimal.



Observations (2/3)

- It should be possible to threshold these values to distinguish between human and machine translations. Note range of H2 values for 201-204 v. 209-213. Threshold might be in the 13-15 range.
- Although recommended that the H2 test be performed on the whole text, it was instead performed on same output segments as the H1 test was performed on with interesting results.



Observations (3/3)

- In both cases, the training effect was unavoidable but mitigated by the level of output difficulty.
- There appears to be some correlation with the results of the other H tests. McCarthy's group had 107 and 207 as the (or one of the) "worst" and 202 as one of the best in the measures they tested. These results correspond with this data.
- These results represent the determinations of one evaluator only, a single data point.