

# **MT Eval Workshop LREC'02**

Report by Eva Forsbom,  
evafo@stp.ling.uu.se,  
Uppsala University,  
Department of Linguistics

# Human-Based Metric

## Correctness/Adequacy/Fidelity

In this evaluation, a human evaluator grades the adequacy of translated segments as compared to a reference translation according to the following scale:

1. None of the meaning expressed in the source fragment is expressed in the translation fragment
2. Little of the source fragment meaning is expressed in the translation fragment
3. Much of the source fragment meaning is expressed in the translation fragment
4. Most of the source fragment meaning is expressed in the translation fragment
5. All meaning expressed in the source fragment appears in the translation fragment

# Human-Based Metric

## Strong Point

- Inter-evaluator consistency – similar rankings.

## Weak Points

- Intra-evaluator consistency – drifting grading level when several translations of the same text.
- Inter-evaluator consistency – different grading levels.
- Interference/Compensation – syntax, stylistics, and spelling affect the grading, by over- or undercompensated grades.
- Quality of reference translation – “verbatim” reference translations tend to penalise freer, often better, translations.
- Segment length – longer segments tend to get lower grades.
- Alternative translations – non-disambiguated translations tend to get higher grades.

# Human-Based Metric

## Results – Score by Evaluator, 1

Transl	Eval1	Eval2	Eval3	Eval4	Total
101	57.5	58.8	41.2	60.0	41.4
102	48.8	55.0	38.8	55.0	37.6
103	47.5	56.2	35.0	61.3	38.1
104	55.0	57.5	43.8	61.3	41.4
105	52.5	53.8	31.2	56.2	36.9
106	52.5	56.2	32.5	58.8	38.1
107	35.0	40.0	25.0	61.3	30.7
108	46.2	52.5	28.7	60.0	35.7
109	47.5	55.0	28.7	61.3	36.7
110	42.5	53.8	32.5	61.3	36.2
111	40.0	55.0	30.0	61.3	35.5
112	37.5	52.5	30.0	58.8	34.0
113	35.0	55.0	35.0	61.3	35.5

# Human-Based Metric

## Results – Score by Evaluator, 2

Transl	Eval1	Eval2	Eval3	Eval4	Total
201	23.8	27.6	18.1	24.8	30.9
202	25.7	33.3	21.9	35.2	38.1
203	32.4	37.1	31.4	37.1	45.3
204	28.6	34.3	22.9	36.2	40.0
205					
206					
207	15.2	21.0	14.3	27.6	25.6
208	18.1	33.3	17.1	35.2	34.1
209	18.1	30.5	16.2	33.3	32.2
210	23.8	32.4	18.1	35.2	35.9
211	20.0	34.3	18.1	35.2	35.3
212	16.2	24.8	14.3	32.4	28.7
213	22.9	32.4	16.2	35.2	35.0

# Human-Based Metric

## Results – Rank by Evaluator, 1

Transl	Eval1	Eval2	Eval3	Eval4	Total
101	1	1	2	8	1
102	5	5	3	13	5
103	6	3	4	1	3
104	2	2	1	1	1
105	3	9	8	12	6
106	3	3	6	10	3
107	12	13	13	1	13
108	8	11	11	8	9
109	6	5	11	1	7
110	9	9	6	1	8
111	10	5	9	1	10
112	11	11	9	10	12
113	12	5	4	1	10

# Human-Based Metric

## Results – Rank by Evaluator, 2

Transl	Eval1	Eval2	Eval3	Eval4	Total
201	4	9	4	11	9
202	3	4	3	3	3
203	1	1	1	1	1
204	2	2	2	2	2
205					
206					
207	11	11	10	10	11
208	8	4	7	3	7
209	8	8	8	8	8
210	4	6	4	3	4
211	7	2	4	3	5
212	10	10	10	9	10
213	6	6	8	3	6

# Automated Metric

## Named Entity Translations

In this evaluation, some human annotators marks up named entities (NE) in a reference translation. All unique NE's from the reference translation are then searched in the translations, and all unique occurrences counted. Some normalisation processes could also be applied:

- No normaliation (NONE)
- Case folding (CASE)
- Diacritica to non-diacritica conversion (DIA)
- Number normalisation (NUMB)
- Removal of possessives (no occurrence)
- Combinations (CASE&DIA, ..., ALL)



# Automated Metric

## Strong Point

- Measures NE translations.

## Weak Points

- Only relevant when many NE's.
- Diacritica conversion tends to boost bad translations (e.g. *Émirate*)
- Quality of reference translation matters (e.g. 60 %).
- Does not seem to correlate with the adequacy metric, not even before normalisation.

# Automated Metric

## Results – Score by Norm. Type, 1

Transl	NONE	NUMB	DIA	CASE&DIA	ALL
Ref	100.0	100.0	100.0	100.0	100.0
101	0.0	0.0	0.0	0.0	0.0
102	0.0	0.0	0.0	0.0	0.0
103	0.0	0.0	0.0	0.0	0.0
104	0.0	0.0	0.0	0.0	0.0
105	0.0	0.0	0.0	0.0	0.0
106	0.0	0.0	0.0	0.0	0.0
107	0.0	0.0	0.0	0.0	0.0
108	0.0	0.0	0.0	0.0	0.0
109	0.0	0.0	0.0	0.0	0.0
110	0.0	0.0	0.0	0.0	0.0
111	0.0	0.0	0.0	0.0	0.0
112	0.0	0.0	0.0	0.0	0.0
113	0.0	0.0	0.0	0.0	0.0

# Automated Metric

## Results – Score By Norm. Type, 2

Transl	NONE	NUMB	DIA	CASE&DIA	ALL
Ref	100.0	100.0	100.0	100.0	100.0
201	50.0	50.0	50.0	50.0	50.0
202	100.0	100.0	100.0	100.0	100.0
203	87.5	87.5	87.5	87.5	87.5
204	50.0	50.0	50.0	50.0	50.0
205					
206					
207	75.0	87.5	75.0	87.5	100.0
208	62.5	62.5	62.5	62.5	62.5
209	50.0	50.0	62.5	50.0	62.5
210	62.5	75.0	62.5	62.5	75.0
211	62.5	75.0	62.5	62.5	75.0
212	87.5	100.0	87.5	87.5	100.0
213	62.5	75.0	62.5	62.5	75.0

# Automated Metric

## Results – Rank by Norm. Type, 1

Transl	NONE	NUMB	DIA	CASE&DIA	ALL
101	1	1	1	1	1
102	1	1	1	1	1
103	1	1	1	1	1
104	1	1	1	1	1
105	1	1	1	1	1
106	1	1	1	1	1
107	1	1	1	1	1
108	1	1	1	1	1
109	1	1	1	1	1
110	1	1	1	1	1
111	1	1	1	1	1
112	1	1	1	1	1
113	1	1	1	1	1

# Automated Metric

## Results – Rank By Norm. Type, 2

Transl	NONE	NUMB	DIA	CASE&DIA	ALL
201	9	9	10	9	10
202	1	1	1	1	1
203	2	3	2	2	4
204	9	9	10	9	10
205					
206					
207	4	3	4	2	1
208	5	8	5	5	8
209	9	9	5	9	8
210	5	5	5	5	5
211	5	5	5	5	5
212	2	1	2	2	1
213	5	5	5	5	5