Correctness of machine translation: A machine translation post-editing task

Maarit Koponen, University of Helsinki

3rd MOLTO Project Meeting – Open Day
Helsinki, September 1, 2011

Machine translation quality and purpose

- Quality standards for human translations are set very high.
  - Canadian official translation evaluation system Sical demands 0 errors for publication quality.
  - Evaluation systems for translator training and qualification often allow a small number of minor errors or one serious error.
- Should we expect the same from machine translations?
- What is the purpose of machine translation?
  - Post-editing
  - Gisting

➔Even a translation with multiple errors is good enough if the reader/translator can interpret the meaning and edit as needed.

Assessing quality with a post-editing task

- A post-editing task was suggested by Philipp Koehn (NAACL HLT 2010) and adopted for the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR 2010.
- Test subjects post-edit raw machine translations without access to the source text.
- The post-edited versions are then evaluated for acceptability with a strict standard of correctness: a fluent translation that contains the same meaning in the document context.
  - Acceptability varied from 26% to 35% (Koehn 2010) and 10% to 80% (Callison-Burch et al. 2010).
  - In Koehn (2010), human translations achieved only 60% acceptability!
- But are sentences ranked unacceptable due to language or meaning?

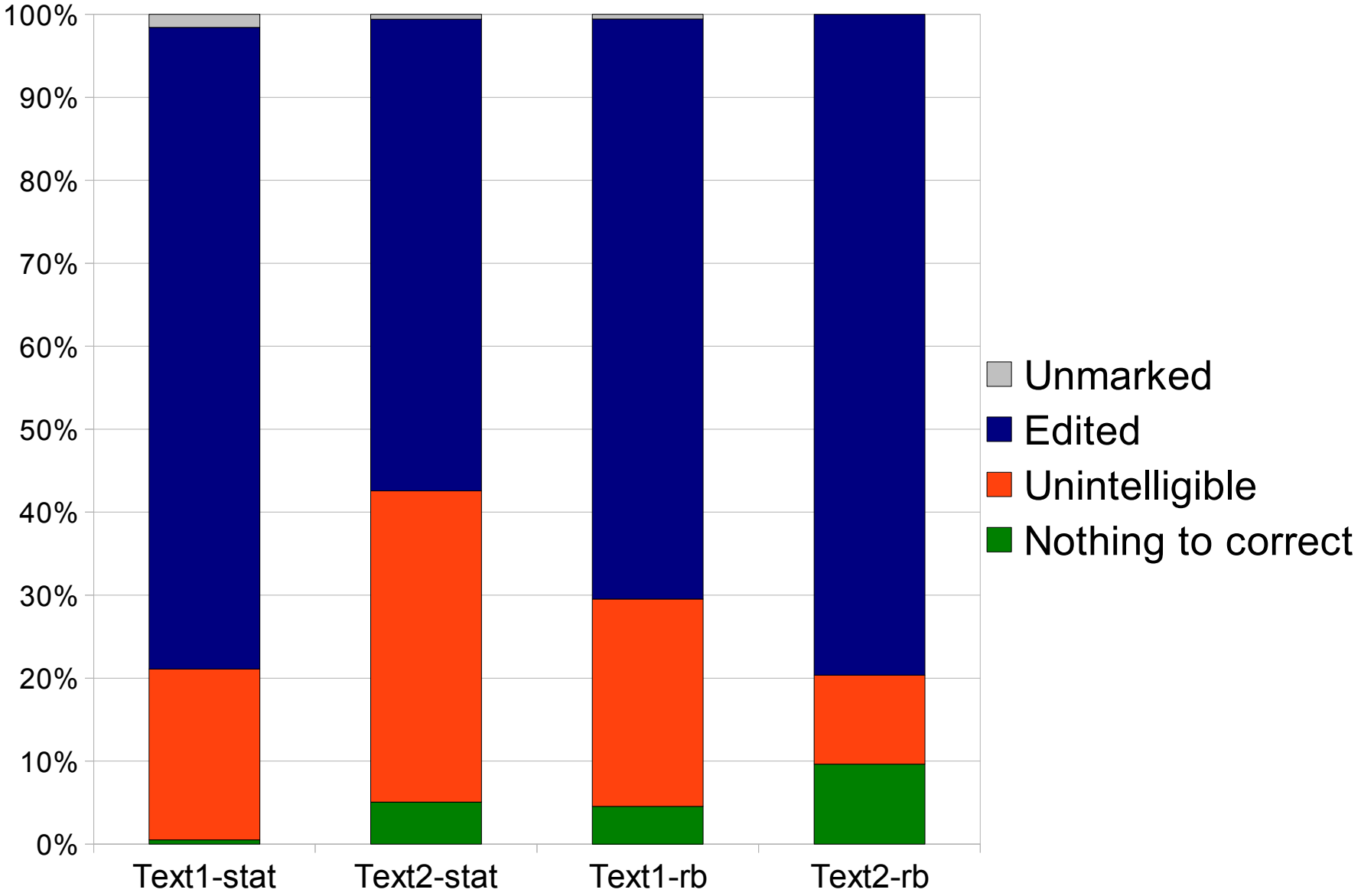
Post-editing task: Study setup

- A course assignment for translator students in an introductory translation technology course at University of Turku, collaboration with Leena Salmi at U of T (Jan 2011).
  - Additional versions from students at the U of H (Spring 2011).
- Two English newspaper articles (~700 words each) were machine translated into Finnish using two systems (statistical Google Translator and rule-based Sunda).
- Test subjects were instructed to edit the text (based on raw MT only) into fluent and clear Finnish according to how they interpret the meaning.
  - ”Nothing to correct” if they felt no editing was needed.
  - ”Unintelligible” if they felt unable to edit at all.

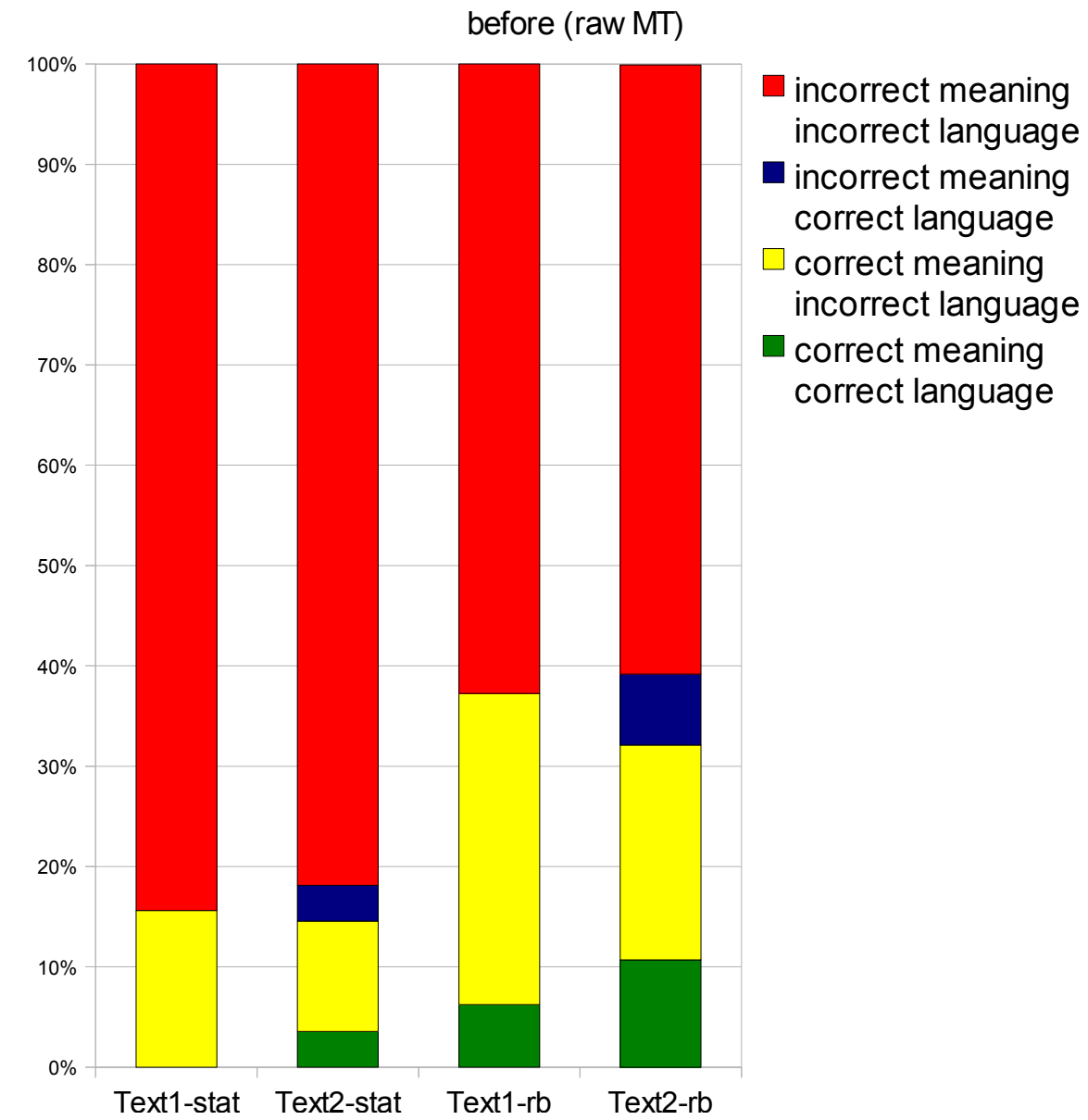
Evaluation of correctness (raw MT and post-edited sentences)

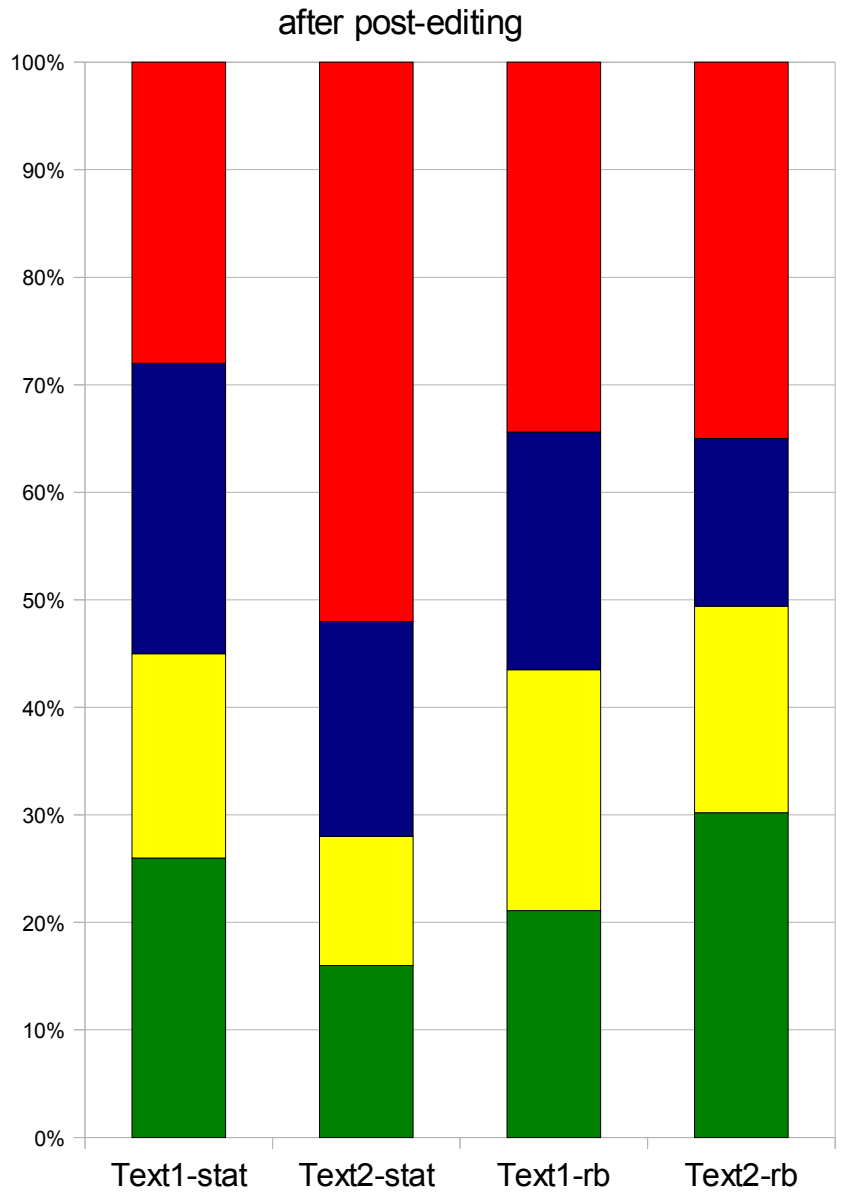
- Correctness was evaluated on a sentence-by-sentence basis.
- Correctness of meaning (compared to source text) and correctness of language (compared to target language conventions) were evaluated separately.
  - Correct meaning – correct language
  - Correct meaning – incorrect language
  - Incorrect meaning – correct language
  - Incorrect meaning – incorrect language
- The two authors first conducted the evaluation separately.
  - Agreement in 65% to 70% of sentences.
  - Evaluators discussed differing cases and agreed on final evaluation.

Edited and unedited sentences



Correctness before and after post-editing





When is post-editing easy?

Example 1

ST: But that won't stop scientists like Barclay from trying to give his new chums a **proper name** – that is to say, **a Latin one**.

MT:Mutta se ei estä tiedemiehiä kuten Barclaytä yrittämästä antaa hänen uusille kavereillensa **erisnimen** – toisin sanoen, **latinalainen**.

'But that won't stop scientists like Barclay from trying to give his new chums a **proper noun** – that is to say, **Latin**.'

(Text2-rb – 10/11 CM, 1/11 IM)

Example 2

ST: So I was surprised to be confronted by **an unidentifiable species while having a sandwich** in the museum's garden," Barclay says.

MT:Joten minä hämmästyin kohdatessani **tunnistamattomat lajit ollessa voileipä** museon puutarhassa", Barclay sanoo.

'So I was surprised to confront the **unidentified (pl.) species (pl.) while being a sandwich** in the museum's garden," Barclay says.'

(Text2-rb – 9/11 CM, 2/11 IM)

When is post-editing not even attempted?

Example 3

ST: Correspondence with **colleagues around Europe** led Barclay to discover that the insect, which resembles the **common** North American **box elder bug**, is actually **most** closely **related** to Arocatus roeselii, a relatively rare species **of seed eaters** usually found in central Europe.

MT: Kirjeenvaihto **kollegoiden Euroopassa johti Barclay huomaamaan**, että hyönteinen, joka muistuttaa **yhteistä** Pohjois-Amerik**an laatikko vanhemman vika**, on itse asiassa **suurin osa liittyy** läheisesti Arocatus roeselii, suhteellisen harvinainen **laji siementen syöjät** yleensä **löytyy** Keski-Euroopassa.

'Correspondence **in the colleagues' Europe led Barclay discover**, that the insect, which resembles a **shared** North America**'s box the elder's fault**, is actually **the largest part** is closely **associated** Arocatus roeselii, relatively rare **species eaters of seeds** generally **is found** in Central Europe.'

(Text2-stat – 9/12 “unintelligible”, 3/12 IM)

When does post-editing fail?

Example 3b

ST: **Correspondence** with colleagues around Europe led Barclay to discover that the insect, which resembles the common North American **box elder bug**, is actually **most closely related** to Arocatus roeselii, a relatively rare species of seed eaters usually **found** in central Europe.

MT: **Vastaavuus** kollegojen välillä ympäri Eurooppaa sai Barclayn havaitsemaan, että**,** hyönteinen, joka muistuttaa tavallista pohjoisamerikkalaista **laatikkoa**, **vanhempi vika**, on itse asiassa**, läheisesti liittyi Arocatusin roeselii?**, siemenensyöjien suhteellisen harvinaiset lajit tavallisesti **löysivät** keskeisestä Euroopasta.

'**Equivalence** between colleagues around Europe led Barclay to discover that, the insect, which resembles an ordinary North American **box**, **an older defect,** is actually, **was associated** closely Arocatus**'s** roeselii**?**, relatively rare species of seed eaters usually **they found** in central Europe.'

(Text2-rb – 4/11 “unintelligible”, 1/11 “nothing to correct”, 6/11 IM)

When does post-editing fail?

Example 4

ST: Her research in Switzerland and France found that even when **people are given unlimited cheap or free calls**, **the number and length of calls does not increase** significantly.

MT: Hänen tutkimuksensa Sveitsissä ja Ranskassa havaitsi, että jopa silloin kun **ihmisille soitetaan**, **ei kasva** merkittävästi.

'Her research in Switzerland and France found that even when **people are called**, **does not increase** significantly.'

(Text1-rb – 7/12 “unintelligible”, 5/12 IM)

Example 5

ST: Barclay is not convinced that climate change is **responsible for Britain’s new inhabitants**.

MT: Barclay ei ole vakuuttunut siitä, että ilmastonmuutos on **vastuussa Britannian asukkaille**.

'Barclay is not convinced that climate change is **accountable to Britain's inhabitants**.'

(Text2-stat – 6/12 “unintelligible”, 1/12 CM, 5/12 IM)

Conclusions

- Meaning was understood and edited correctly in 26% to 49% of sentences.
- Language errors were often ignored by test subjects.
- Willingness to edit and success rate varied greatly between test subjects.
- Sometimes recovering from errors is easy.
  - Meaning can be deduced from context and general knowledge.
- But other errors affect meaning in critical ways.
  - Multiple errors affect long passages.
  - Key piece of information not deducible from context or general knowledge is missing or garbled.
  - Some errors are not even evident!

Literature

Bensoussan, Marsha & Judith Rosenhouse (1990). “Evaluating student translations by discourse analysis.” *Babel,* 36(2), 65-84.

Callison-Burch, Chris et al. (2010). “Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation.” *ACL 2010: Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Proceedings of the workshop,* 17-53.

Koehn, Philipp. (2010). “Enabling monolingual translators: Post-editing vs. options.” *NAACL HLT 2010: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. *Proceedings,* 537-545.

Martínez Melis, Nicole & Amparo Hurtado Albir (2001). “Assessment in Translation Studies: Research Needs.” *Meta* 46(2), 272–287.

Penttilä, Ari (2008). “Seeking an optimal system for certifying translators. Finnish experiences over the past 40 years.” *Proceedings of the XVIII FIT World Congress*. Publication on CD-ROM, no page numbers.

Secară, Alina (2005). “Translation Evaluation – a State of the Art Survey.” *Proceedings of the eCoLoRe/MeLLANGE Workshop*, Leeds, 21-23 March 2005, 39–44. On line at: http://www.leeds.ac.uk/cts/research/publications/leeds-cts-2005-03-secara.pdf (consulted 30.8.2011)

Williams, Malcolm (2001). “The Application of Argumentation Theory to Translation Quality Assessment.” *Meta* 46(2), 326–344.