# Chunk Accuracy: A Simple, Flexible Metric for Translation Quality

Lars Ahrenberg

Department of Computer and Information Science
Linköping University

MTE workshop, Reykyavik
May 26, 2014

# Overview

Introduction

Error-based approaches

The proposal

# Translation Quality Assessment through metrics

- Equivalence metrics e.g. Tarvi, 2004

# Translation Quality Assessment through metrics

- Equivalence metrics e.g. Tarvi, 2004
- Criteria-based
  - tabulated (yes, no, N/A)

# Translation Quality Assessment through metrics

- Equivalence metrics e.g. Tarvi, 2004
- Criteria-based
  - tabulated (yes, no, N/A)
- Error-based
  - error count
  - error profile

# Translation Quality Assessment through metrics

- Equivalence metrics e.g. Tarvi, 2004
- Criteria-based
  - tabulated (yes, no, N/A)
- Error-based
  - error count
  - error profile
- Reference-based
  - a single metric, or
  - a set of metrics

# Problems with reference-based metrics

They are indirect!

# Problems with reference-based metrics

They are indirect!

- quality is usually not assessed

# Problems with reference-based metrics



They are indirect!

- quality is usually not assessed
- purpose of reference may differ from system translations

# Problems with reference-based metrics



They are indirect!

- quality is usually not assessed
- purpose of reference may differ from system translations
- references cannot cover the space of possible translations

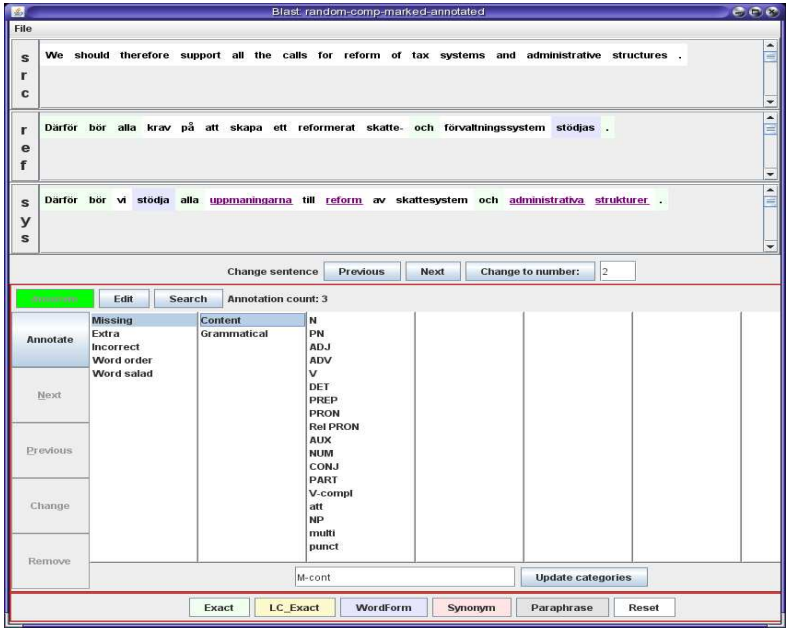
# Overview

Introduction

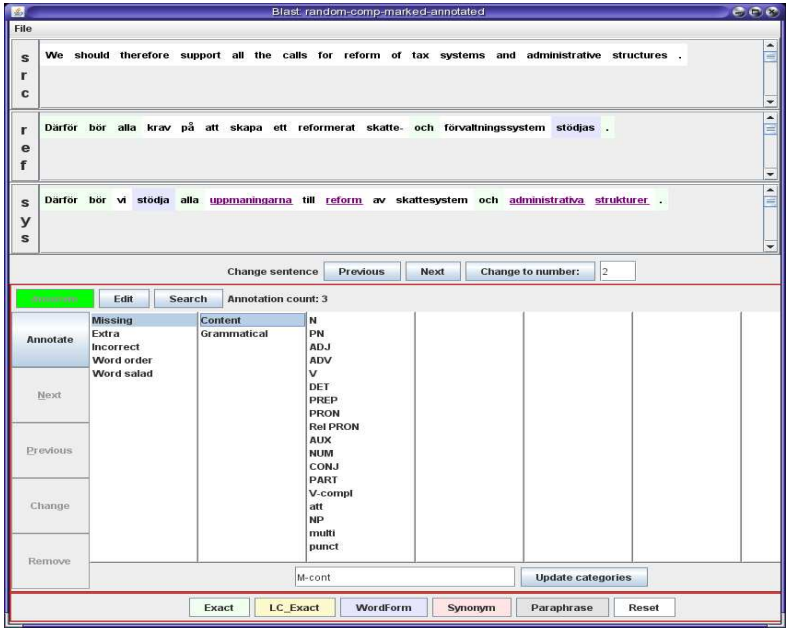
Error-based approaches

The proposal

# Lessons from Stymne & Ahrenberg, 2012

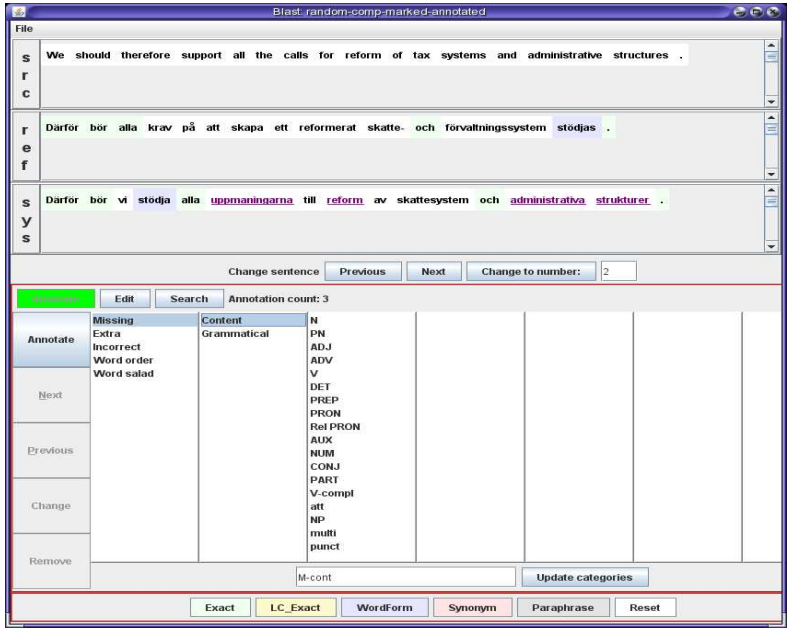


# Lessons from Stymne & Ahrenberg, 2012



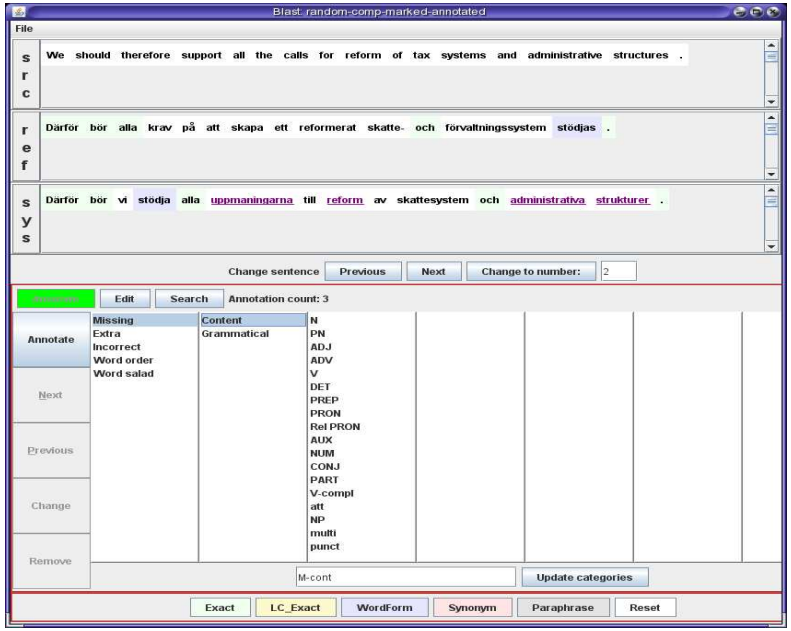

- disagreement on which token was affected

# Lessons from Stymne & Ahrenberg, 2012



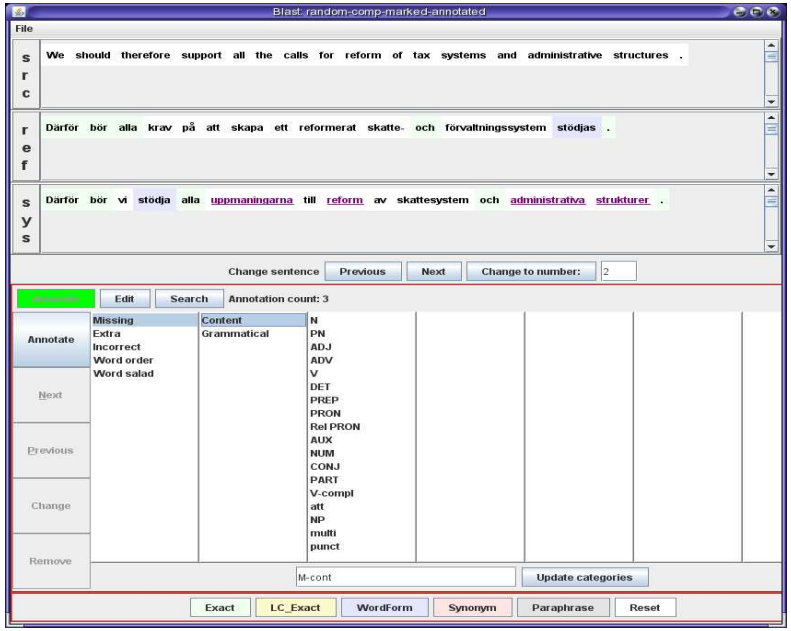

- disagreement on which token was affected
- disagreement on whether error is present

# Lessons from Stymne & Ahrenberg, 2012




- disagreement on which token was affected
- disagreement on whether error is present
- the more detail the less agreement
  (cf. Bayerl & Paul, 2011)

# Lessons from Stymne & Ahrenberg, 2012




- disagreement on which token was affected
- disagreement on whether error is present
- the more detail the less agreement
  (cf. Bayerl & Paul, 2011)
- short sentences didn't represent the full set well

# Quality aspects

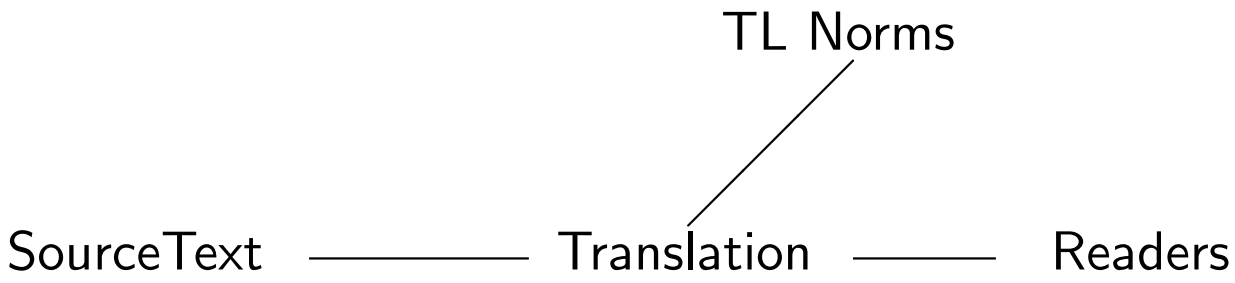
Translation

# Quality aspects

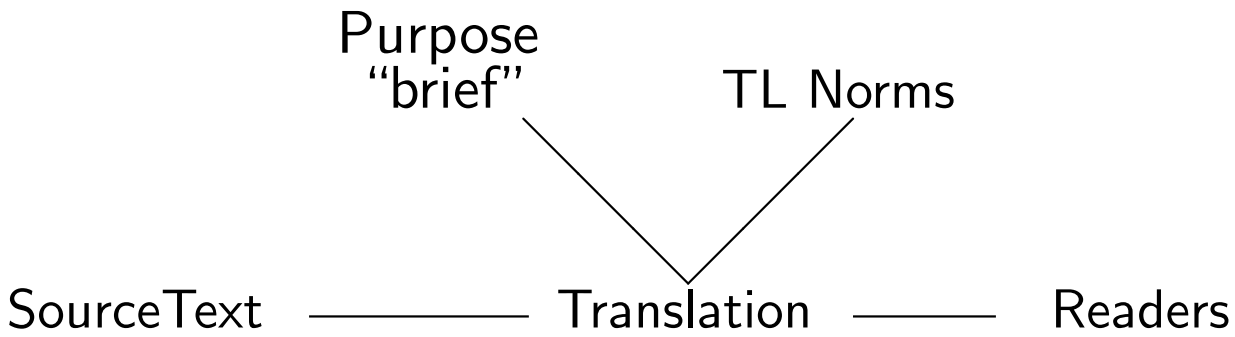
SourceText ——— Translation

# Quality aspects

# Quality aspects



# Quality aspects



# General problems with error-based metrics

- the error taxonomy is not selected in relation to purpose

# General problems with error-based metrics

- the error taxonomy is not selected in relation to purpose
- inter-annotator agreement tends to be low

# General problems with error-based metrics

- the error taxonomy is not selected in relation to purpose
- inter-annotator agreement tends to be low
  - errors may often be corrected (and thus classified) in many ways

# General problems with error-based metrics

- the error taxonomy is not selected in relation to purpose
- inter-annotator agreement tends to be low
  - errors may often be corrected (and thus classified) in many ways
  - large, detailed taxonomies are useless for really bad translations

# General problems with error-based metrics

- the error taxonomy is not selected in relation to purpose
- inter-annotator agreement tends to be low
  - errors may often be corrected (and thus classified) in many ways
  - large, detailed taxonomies are useless for really bad translations
  - boring task

# General problems with error-based metrics

- the error taxonomy is not selected in relation to purpose
- inter-annotator agreement tends to be low
  - errors may often be corrected (and thus classified) in many ways
  - large, detailed taxonomies are useless for really bad translations
  - boring task
- global (macro-level) properties are not considered

# General problems with error-based metrics

- the error taxonomy is not selected in relation to purpose
- inter-annotator agreement tends to be low
  - errors may often be corrected (and thus classified) in many ways
  - large, detailed taxonomies are useless for really bad translations
  - boring task
- global (macro-level) properties are not considered
- it is not obvious how to define informative metrics (cf. Williams, 2001)

# Overview

Introduction

Error-based approaches

The proposal

# Main features

- Allocate issues to chunks – instead of words

# Main features

- Allocate issues to chunks – instead of words
- Make the taxonomy small and relevant

# Main features

- Allocate issues to chunks – instead of words
- Make the taxonomy small and relevant
- Each issue type should contribute to the translation profile

# Example

- *These star-forming region, called by scientists unromantisch*

# Example: agreement issues

- **These** *star-forming* **region**, ...

# Example: word order issue

Word order uncertainties

- *..., called by scientists unromantisch*

# Taxonomies

Example: post-editing

- Good
- Editable
- Useless

# Taxonomies

Example: assimilation / reading support

- Comprehensible wihout effort
- Comprehensible with effort
- Incomprehensible

# Metrics

- A profile: the percentage of chunks falling into any given category
- A single metric: a weighted combination of these percentages

# Example



| **Category** | **Weight** | **Result** |
|---|---|---|
| Comprehensible | 1.0 | 0.40 |
| Compr. with effort | 0.5 | 0.25 |
| Incomprehensible | 0 | 0.35 |

Single metric: 0.40 + 0.5*0.25 = 0.525

# Chunk selection

# Chunk selection

- a single clause, or a long phrase

# Chunk selection

- a single clause, or a long phrase
- a discourse unit (?)

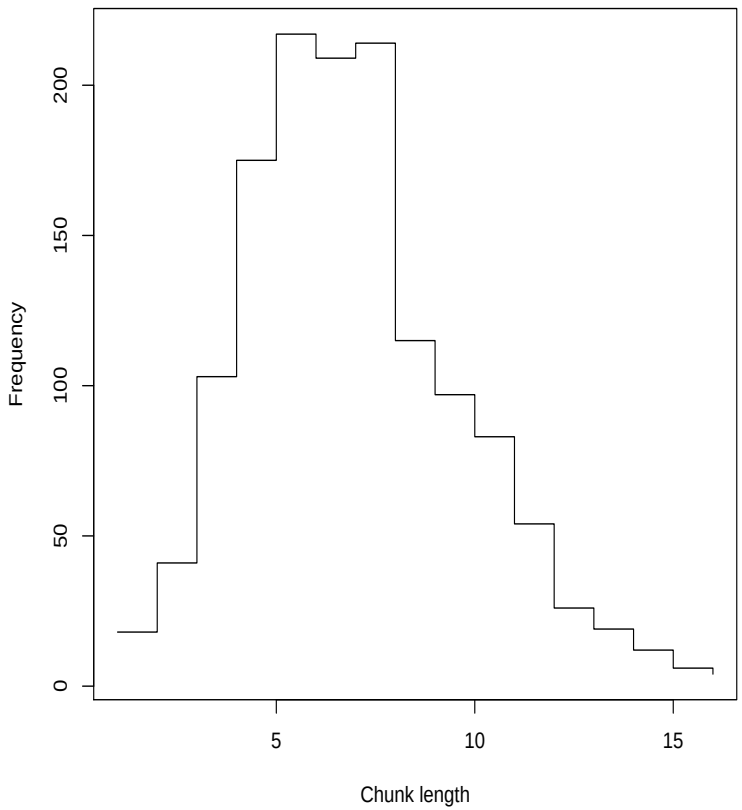
# Chunk selection

- a single clause, or a long phrase
- a discourse unit (?)
- based on source chunks (?)

# Chunk selection

- a single clause, or a long phrase
- a discourse unit (?)
- based on source chunks (?)
- goal: automatic

# Chunk lengths - manual sample




mean: 6.5 words

# The End

Thank you!