Towards the Calibration of Translation Errors: the Why and the How

June Eyckmans, Philippe Anckaert & Winibert Segers
What about translation assessment literacy within a European context?

• Who is looking out for test quality?
• To quote Charles Alderson:
  “Who takes on board the responsibility of test quality control?”
• Special Interest Group on Translation and Interpretation Assessment
• After Q&A: introduction of the SIG
Starting out with a paradox ...

• Lado (1961)
  “ironically, translation tests that are so common in testing other skills are not available as tests of the ability to translate”

• Today: no development of tests for measuring translation ability that allow psychometric control of the reliability of these measures
Why?

- Two explanations:
  1) lack of validity of the translation test as measure for language proficiency (Klein-Braley, 1987) → loss of popularity of the format
  2) “corporate culture” among translation trainers/language teachers → reticent about the use of psychometrics → epistemological problem
“it seems unlikely that translation quality assessment can ever be objectified in the manner of natural science” (House 1981:64)
Educational context

• Today’s practice of evaluating students’ translations: characterized by the use of assessment sheets/analytical grids

• Taxonomy of mistakes / bonuses
  → near exhaustive identification of different kinds of mistakes
  → relative “weight” of the mistake

= criterion-related approach
Why analytical grids?

→ to enhance reliability of the evaluation

Factors that threaten reliability:
- number of translations to be scored
- time pressure
- order of correction: contrast effect
- halo effect: unconscious preconceptions about students with a weak/strong reputation
- personal views on the nature/essence of translation ability
Analytical grids

- Useful because:
  - criteria reflect points of interest of the evaluators
  - valuable consensus on what to accept/reject within an organization
- Problematic because:
  - difficult to operationalize consistently (doubtful stability of the criteria)
  - essentially subjective approach (cases of dispute: lack of justification \(\rightarrow\) litigations)
Professional context

- Use of matrices
  (approach similar to analytical grids)
- Discussions on evaluation focuses on customer-related service
SAE J2450
(Society of Automotive Engineers)

Eight categories
1) wrong term
2) omission
3) grammatical error related to word structure, agreement and part of speech
4) wrong word order
5) misspelling
6) punctuation error
7) superfluous text
8) miscellaneous errors
In short

Both in educational and professional context

→ attempts at defying a subjective, “impressionistic” evaluation

→ analytical grids / matrices

= setting up a taxonomy of categories in an attempt to grasp the mental attributes that constitute translating ability
Translation competence

• Invisible – cannot be observed directly
• Translation skills do not equal language skills
EN 15038

• Five-fold description
  1) translation competence
  2) source/target language competence
  3) research competence
  4) cultural competence
  5) technical competence
EN 15038

Definition of competences
→ demand for certificates
→ How to determine that a candidate corresponds to this five-fold profile?
→ Need for reliable descriptors of competence
Descriptors of competence

➔ Evaluation will be performance-based
➔ Method to relate performance indicators to the underlying translation competence
Questions to be dealt with ... 

• What kind of performance is indicative of the construct translation ability?
• Can/should the translation competence be split up into distinguishable subcomponents?
• If so,
  -should these be measured independently?
  -chronological acquisition pattern of main skills and subskills? (empirical evidence?)
Proposal

• Opt for a global evaluation encompassing all aspects of translation ability.

For:
- in actual performances, subcomponents are more or less inextractable
- mistakes as well as bonuses originate from the interaction of a particular text with a particular translator
Norm-related approach

= by means of a representative sample survey mistakes are identified that are shown to have discriminating power.

⇒ method of evaluation that is independent of subjective a priori notions
How?

- Calibration of Dichotomous Items (CDI)
- Transfer the “item”-concept to translation practice
- Items = selected on the basis of a pre-test
- Use: summative evaluation (on which to base high stake decisions)
Prerequisites

- Dichotomous approach
- Inevitable pre-testing (sampling)
Procedure

1. candidate translators translate a text
2. these translation performances are corrected and all mistakes are registered (cumulative)
3. resulting in a total number of potential “items”
4. potential “items” receive 0/1 value in matrix
5. determine “Corrected Item-Total Correlation” values
6. estimate reliability (Cronbach’s Alpha)
7. select items with high “Corrected Item-Total Correlation” values (> .30) for inclusion in a calibrated translation test
<table>
<thead>
<tr>
<th>Media</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>en amusement</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>trekken alle aandacht naar zich toe</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Een van de grootste misvattingen</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>die ik bij de meeste mensen heb vastgesteld</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>over reclame,</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>is dat zij ervan uitgaan dat</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>al het geïnvesteerde reclamegeld</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Media en amusement trekken alle aandacht naar zich toe
Een van de grootste misvattingen die ik bij de meeste mensen heb vastgesteld 
/over reclame,/ is dat zij ervan uitgaan dat/ al het geïnvesteerde 
reclamegeld/ van bedrijven terechtkomt in de 'reclamewereld', /waarmee zij 
de wereld van reclamebureaus en reclamemakers bedoelen/. Dat is niet zo. 
/Het overgrote deel van de reclame-investeringen/ van het bedrijfsleven 
gaat naar de aankoop van ruimte in de media, en komt dus terecht /op de 
bankrekeningen/ van de mediasubjectpen met hun tijdschriften, kranten, 
radiozenders, /tv-stations/, bioscopen, /billboards/... Bedrijven zijn immers 
op zoek naar een publiek om hun producten /bekend/ en geliefd te maken, in 
de hoop dat publiek ervan te kunnen overtuigen hun producten ten minste eens 
te proberen. /Dat publiek/ wordt geleverd door de media. De bedrijven kopen 
dus pagina's of zendtijd, /{qu’ils}/ en kunnen zo in contact treden met het 
publiek /van die media/. /Op die manier ontstaat/ er een miljardenstroom 
van reclamegeld (in België meer dan 1,75 miljard euro per jaar), /die van de 
bedrijven/naar de /media/ stroomt.
Calibrated Translation Test

• When the test is administered, only the calibrated items need to be corrected.
  >> univocal results, regardless of the evaluator
  >> time-saving procedure
• Safety check: item calibration needs to be checked on stability
• Also: test robustness needs to be explored. Tests will have to be validated for different language combinations and language uses/registers.
Constraints of the method

- representative nature of the sample is of overriding importance
- implementation of the method + construction of a reliable battery of tests \(\rightarrow\) constant monitoring
- sufficiently large populations (to safeguard reliability) \(\rightarrow\) cooperation recommendable
- colleagues initial reticence with regard to a psychometric approach of evaluation
Advantages of the method

-use of the discriminating power of items → an evaluation that is reliable and much more precise
-flaws inherent in the evaluator or the text do not undermine the method → stable and evaluator-independent evaluation
-the method is inclusive of every possible dimension of translation ability
Putting our claims to the test in a controlled experiment

Goals:
1) Compare holistic, analytic and CDI scoring method in terms of reliability
2) Construct standardized test for Dutch ➔ French translation
Empirical data

• Task: translation of two 300 word texts (equivalent) Dutch → French
• Participants: 124 participants, all students within a four-year translator training programme (BA1 → MA)
• Corrected by experienced translators and revisors according to holistic or analytic method (+ CDI-method)
Profiles assessors

• Two translator trainers + two revisors
• Trained as translators, >25 years experience
• Choice of correction method
Scoring method

- Holistic: score based on global impression of translation quality
- Analytic score based on use of matrix (taxonomy of mistakes + weight)
Crossed experimental design

<table>
<thead>
<tr>
<th>Text A / Group 1 / BA1</th>
<th>holistic</th>
<th>Text B / Group 2 / BA1</th>
<th>analytic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text A / Group 1 / BA2</td>
<td>holistic</td>
<td>Text B / Group 2 / BA2</td>
<td>analytic</td>
</tr>
<tr>
<td>Text A / Group 1 / BA3</td>
<td>holistic</td>
<td>Text B / Group 2 / BA3</td>
<td>analytic</td>
</tr>
<tr>
<td>Text A / Group 1 / MA</td>
<td>holistic</td>
<td>Text B / Group 2 / MA</td>
<td>analytic</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text B / Group 1 / BA1</th>
<th>holistic</th>
<th>Text A / Group 2 / BA1</th>
<th>analytic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text B / Group 1 / BA2</td>
<td>holistic</td>
<td>Text A / Group 2 / BA2</td>
<td>analytic</td>
</tr>
<tr>
<td>Text B / Group 1 / BA3</td>
<td>holistic</td>
<td>Text A / Group 2 / BA3</td>
<td>analytic</td>
</tr>
<tr>
<td>Text B / Group 1 / MA</td>
<td>holistic</td>
<td>Text A / Group 2 / MA</td>
<td>analytic</td>
</tr>
</tbody>
</table>
Research hypotheses

1) The inter-rater reliability between the evaluators of the holistic and the analytic evaluation method is unconvincing.
   - rank orders will vary
   - unreliable scoring method
2) The rankings of the students’ performances when corrected by the holistic and the analytic method differ according to the text that had to be translated.
   - measurement of ability dependent on text
Preliminary results

- Lack of inter-rater reliability (r = .670) for both methods. Cronbach’s alpha CDI = .94 (k=50, n=63).
- Factors profession and method:
  - revisors attribute higher scores than trainers
  - holistic method yields higher scores than analytic method
- Text is shown to be significant.
Conclusion

-bridges the gap between CLT and the specific epistemological characteristics of translation studies
-reproducible method that relates performance indicators to the underlying translation competence in a psychometrically controlled way (evidence of test quality)
-high stakes test development: battery of tests for different languages and text types
-possibility of measuring the construct translation competence (independently of text)
References


Thank you!