Translation assessment within a common European framework?

June Eyckmans, Philippe Anckaert & Winibert Segers
Introduction

- EALTA 2008 Athens: conference on language assessment literacy → what about translation assessment literacy?
Introduction

• Today:
  1) state of affairs in translation assessment
  2) new method of translation assessment
  3) results controlled empirical design focusing on
     (a) reliability and validity of three assessment methods,
     (b) text independent measurement of translation competence
  4) conclusion with reference to the general theme of this conference (=European collaboration)
State of affairs

• Two striking observations:
  1) notwithstanding the daily practice of translation assessment ➔ not a lot of research
  2) “corporate culture” among translation trainers/language teachers ➔ epistemological problem/denial
“it seems unlikely that translation quality assessment can ever be objectified in the manner of natural science” (House 1981:64)

The preconception that it is impossible to objectify translation quality assessment without surrendering its essence is very tenacious among translator trainers and language teachers.
Current practice

• Educational as well as professional domain → holistic or analytic method.

• Holistic approach: assessment is based on global impression of translation quality (also called intuitive or impressionistic approach)
Analytic method

• Attempt at defying a subjective, “impressionistic” evaluation through the use of analytical grids / matrices

= taxonomy of mistakes + relative “weight” of the mistake
Reliability issues

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
Reliability issues

- Holistic method: essentially subjective approach
- Analytic method: difficult to operationalize consistently (also because of doubtful stability of the criteria)
Translation competence

- Translation skills do not equal language skills
- Several definitions of the construct “translation competence” available
- Since 2006: publication of European Norm
Five-fold description of professional competence that is required:
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?
→ EN itself symptomatic: not a product norm
→ Need for reliable descriptors of competence
→ Within European context: robustness
CDI method

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

Based on the convictions:
- 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
CDI method

= by means of a representative sample survey
text segments with discriminating power are
identified through a process of pre-testing
and item calibration
CDI method

- Transfer the “item”-concept to translation practice
- Use: summative evaluation (on which to base high stake decisions)
- Requires a dichotomous approach and pre-testing (items are selected on the basis of a pre-test)
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 (> .35) for inclusion in a calibrated translation test
Matrix

<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 mediaschepen 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 ➔
  constant monitoring
- sufficiently large populations (to safeguard reliability) ➔ cooperation recommendable
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
Controlled experiment

Aim:
Compare holistic, analytic and CDI scoring method in terms of reliability
Empirical data

• Task: translation of two 300 word texts (equivalent) Dutch ➔ French
• Participants: 95 participants, all students within a four-year translator training programme (BA1 ➔ MA)
• Corrected by experienced translator trainers 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
Crossed experimental design

All groups from all levels translated two texts (A and B). All these performances were scored by four assessors:

- Translator trainer following holistic scoring method
- Translator trainer following analytic scoring method
- Revisor following holistic scoring method
- Revisor following analytic scoring method

All performances are scored according to the CDI method.
Research hypotheses

1) The inter-rater reliability between the evaluators of the holistic and the analytic evaluation method is unconvincing.  

→ rank orders will vary substantially

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
Research hypotheses

3) The reliability of the CDI method will be high.
4) The rankings of the students’ performances when corrected by the CDI method will not differ according to the text that had to be translated.

→ measurement of ability independent of text
Results (1) Inter rater reliability

- Inter-rater reliability is unconvincing:
  between .635 and .780 (A)
  between .742 and .835 (B)
  do not reflect a satisfying inter-rater reliability
  (predictive value very limited)
  countless examples:
    student 94, 0/20, 4/20, 3/20, 14/20
    student 80, 9/20, 18/20, 11/20, 15/20
Results (1)

→ the participants’ chances of succeeding the test would be highly dependent on the grader in question
Results (1)

• Graders of the holistic method obtain a lower agreement on the rank orders of the participants than the graders of the analytic method ($r = .669$ versus $r = .740$)

• ANOVA points to a significant effect of the factors profession and method:
  - revisors attribute higher scores than trainers
  - holistic method yields higher scores than analytic method
Results (2) Text effect

• ANOVA: rankings differ substantially according to the translated text.
• Correlation coeffïënts reveal that only a small amount of the variance is explained:
  teacher analytic: .642 (r² 0.412)
  teacher holistic: .706 (r² 0.499)
  revisor analytic: .712 (r² 0.506) ➔ best case
  revisor holistic: .449 (r² 0.201) ➔ worst case
Text B vs Text A: Teacher with Analytic Method

\[ y = 0.7618x + 19.481 \]

\[ R^2 = 0.4126 \]
Text A vs Text B: Revisor with analytic method

\[ y = 0.732x + 48.313 \]

\[ R^2 = 0.5067 \]
Text A vs Text B: Teacher with holistic method

$y = 0.7768x + 30.561$

$R^2 = 0.499$
Text A vs Texte B: Revisor with holistic method

\[ y = 0.4329x + 52.142 \]

\[ R^2 = 0.2014 \]
Results (2)

• Students would pass or fail depending on the text that would have been used.
Results (3)

- The reliability of the CDI method is high
  Text A, 63 items, Cronbach’s alpha .956
  Text B, 42 items, Cronbach’s alpha .927
Results (4)

• ANOVA shows that the rankings differ substantially according to the translated text in the case of the CDI method as well.
• Correlation coefficiënt .603 ($r^2 0.363$)

→ CDI is reliable and evaluator-independent but not text independent.
Conclusion

-experiment suggests translation assessment highly unreliable, on the basis of our data it is impossible to make a plea for either method

-CDI method:
  1) thanks to matrix scores, item difficulty and item discrimination indices can be calculated, the reliability of the test can be calculated
  2) reproducible method, evidence of test quality
  3) but: we did not succeed in measuring translation competence independent of text

→”relates performance indicators to underlying translation competence in psychometrically controlled way”
Conclusion

Our talk today ties in with a European call for test standardisation in the sense that we are convinced of the need for

→ a reliable and valid certification of translation competence

→ an exploration of test robustness so that tests can be validated for different language combinations

→ collaboration between translation teachers and test developers in order to rise to the methodological challenge of developing a battery of tests for different languages and text types
References


Thank you!
Remarks

-difficult to maintain sound evaluation practice in which you do not only distinguish the really good from the really bad translations but you also discriminate within the large group of mediocre translations

-all scores have to be justifiable and nowadays students exercise their rights on equitable assessment of their performances
Remarks

-practice has shown that assessors do not agree on the exact nature of a mistake or how many point should be substracted or awarded