Mending the Mess Babel Has Left Us With
Cross-Lingual Automatic Coding in International Large-Scale Assessments

Fabian Zehner
DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment (ZIB)
Nov 3, 2022;
2022 MARC Conference
Structure of the Talk

Due to an unfortunate incident, differently than planned ...

Part A: initial evidence from first baseline steps
Part B: conceptual discussion of cross-lingual coding

and ... not most up-to-date references & pagination
Ancestral Traditional Perspective on Text Responses

• more popular in large-scale assessments since 1990s (Bennett, 1993)
• however, viewed largely skeptically (see Bejar, 2017; Bennett & Ward, 1993)
  ◦ marginal gains in construct coverage; e.g. higher-order cognitive skills (Guthrie, 1984)
  ◦ lack of objectivity, reliability, and efficiency...

• Nowadays
  • incremental value in construct validity, i.a., ...
  ◦ computer and information literacy (Ihme, Senkbeil, Goldhammer, & Gerick, 2017)
  ◦ literacy in mathematics (Birenbaum & Tatsuoka, 1987; Bridgeman, 1991)
  ◦ reading literacy (Griffo, 2011; Lim, 2019; Millis, Magliano, Wiemer-Hastings, Todaro, & McNamara, 2011b; Rauch & Hartig, 2010; Rupp, Ferne, & Choi, 2006)
  • sometimes, only certain psychometric properties differ (Schult & Sparfeldt, 2018)
Ancestral Traditional Perspective on Text Responses
Ancestral Traditional Perspective on Text Responses

**Constructed Responses Historically**

- more popular in large-scale assessments since 1990s  
  (Bennett, 1993)
- however, viewed largely skeptically  
  (see Bejar, 2017; Bennett & Ward, 1993)
Ancestral Traditional Perspective on Text Responses

**Constructed Responses Historically**

- more popular in large-scale assessments since **1990s** *(Bennett, 1993)*
- however, viewed largely skeptically *(see Bejar, 2017; Bennett & Ward, 1993)*

  gains in **construct coverage**; e.g. higher-order cognitive skills *(Guthrie, 1984)*

- marginal gains in construct coverage; e.g. higher-order cognitive skills *(Guthrie, 1984)*
- lack of objectivity, reliability, and efficiency...
- Nowadays • incremental value in construct validity, i.a.,... *(Ihme, Senkbeil, Goldhammer, & Gerick, 2017)*
  - computer and information literacy *(Ihme, Senkbeil, Goldhammer, & Gerick, 2017)*
  - literacy in mathematics *(Birenbaum & Tatsuoka, 1987; Bridgeman, 1991)*
  - reading literacy *(Griffo, 2011; Lim, 2019; Millis, Magliano, Wiemer-Hastings, Todaro, & McNamara, 2011b; Rauch & Hartig, 2010; Rupp, Ferne, & Choi, 2006)*
  - sometimes, only certain psychometric properties differ *(Schult & Sparfeldt, 2018)*
Ancestral Traditional Perspective on Text Responses

### Constructed Responses Historically

- **more popular** in large-scale assessments since **1990s**  
  (Bennett, 1993)
- **however, viewed largely skeptically**  
  - marginal gains in **construct coverage**, e.g. higher-order cognitive skills  
    (Guthrie, 1984)  
  - lack of **objectivity, reliability, and efficiency**  
    (see Bejar, 2017; Bennett & Ward, 1993)
Ancestral Traditional Perspective on Text Responses

**Constructed Responses Historically**

- more popular in large-scale assessments since 1990s (Bennett, 1993)
- however, viewed largely skeptically
  - marginal gains in construct coverage; e.g. higher-order cognitive skills (Guthrie, 1984)
  - lack of objectivity, reliability, and efficiency

**...and Nowadays**

- incremental value in construct validity, i.a., ...
  - computer and information literacy (Ihme, Senkbeil, Goldhammer, & Gerick, 2017)
  - literacy in mathematics (Birenbaum & Tatsuoka, 1987; Bridgeman, 1991)
  - reading literacy (Griffo, 2011; Lim, 2019; Millis, Magliano, Wiener-Hastings, Todaro, & McNamara, 2011b; Rauch & Hartig, 2010; Rupp, Ferne, & Choi, 2006)
Ancestral Traditional Perspective on Text Responses

**Constructed Responses Historically**

- more popular in large-scale assessments since 1990s (Bennett, 1993)
- however, viewed largely skeptically (see Bejar, 2017; Bennett & Ward, 1993)
  - marginal gains in construct coverage; e.g. higher-order cognitive skills (Guthrie, 1984)
  - lack of objectivity, reliability, and efficiency

**...and Nowadays**

- incremental value in construct validity, i.a., ...  
  - computer and information literacy (Ihme, Senkeil, Goldhammer, & Gerick, 2017)
  - literacy in mathematics (Birenbaum & Tatsuoka, 1987; Bridgeman, 1991)
  - reading literacy (Griffo, 2011; Lim, 2019; Millis, Magliano, Wiener-Hastings, Todaro, & McNamara, 2011b; Rauch & Hartig, 2010; Rupp, Ferne, & Choi, 2006)
- sometimes, only certain psychometric properties differ (Schult & Sparfeldt, 2018)
Text Responses with Automatic Processing

Natural Language Processing and Machine Learning allow to . . .

- process and classify responses
- access unstructured information for secondary analysis (at the large scale)
Text Responses with Automatic Processing

Natural Language Processing and Machine Learning allow to . . .

- process and classify responses
- access unstructured information for secondary analysis (at the large scale)
- extract information beyond scores (e.g., He, 2013; Zehner, Goldhammer, & Sälzer, 2018)
Text Responses with Automatic Processing

**Natural Language Processing and Machine Learning allow to . . .**

- process and classify responses
- access unstructured information for secondary analysis (at the large scale)
- extract information beyond scores

  (e.g., He, 2013; Zehner, Goldhammer, & Sälzer, 2018)

**Historically, . . .**

- the start: automatic essay grading in 1960s

  (Page, 1966, 1968)
# Text Responses with Automatic Processing

## Natural Language Processing and Machine Learning allow to . . .

- process and classify responses
- access unstructured information for secondary analysis (at the large scale)
- extract information beyond scores

(e.g., He, 2013; Zehner, Goldhammer, & Sälzer, 2018)

## Historically, . . .

- the start: automatic essay grading in 1960s
  (Page, 1966, 1968)
- automatic short answer grading started in 1990s
- countless developments since 2010s, two competitions
  (Hewlett Foundation, 2012; Dzikovska et al., 2013)
- deep learning and—since 2017—transformers in particular
  (e.g., Sung, Dhamecha, & Mukhi, 2019)
- large-scale assessments neglected, 2 exceptions
  (Yamamoto, He, Shin, & von Davier, 2018; Zehner, Sälzer, & Goldhammer, 2016)
Text Responses with Automatic Processing

Natural Language Processing and Machine Learning allow to . . .

- process and classify responses
- access unstructured information for secondary analysis (at the large scale)
- extract information beyond scores

Historically, . . .

- the start: automatic essay grading in 1960s (Page, 1966, 1968)
- countless developments since 2010s, two competitions (Hewlett Foundation, 2012; Dzikovska et al., 2013)

(out-of-date overview at Burrows et al., 2014)
Text Responses with Automatic Processing

Natural Language Processing and Machine Learning allow to . . .

- process and classify responses
- access unstructured information for secondary analysis (at the large scale)
- extract information beyond scores

Historically, . . .

- the start: automatic essay grading in 1960s
  (Page, 1966, 1968)
- automatic short answer grading started in 1990s
- countless developments since 2010s, two competitions
  (Hewlett Foundation, 2012; Dzikovska et al., 2013)
  (out-of-date overview at Burrows et al., 2014)
- deep learning and—since 2017—transformers in particular
  (e.g., Sung, Dhamecha, & Mukhi, 2019)
Text Responses with Automatic Processing

Natural Language Processing and Machine Learning allow to . . .

- process and classify responses
- access unstructured information for secondary analysis (at the large scale)
- extract information beyond scores (e.g., He, 2013; Zehner, Goldhammer, & Sälzer, 2018)

Historically, . . .

- the start: automatic essay grading in 1960s (Page, 1966, 1968)
- countless developments since 2010s, two competitions (Hewlett Foundation, 2012; Dzikovska et al., 2013) (out-of-date overview at Burrows et al., 2014)
- deep learning and—since 2017—transformers in particular (international) (e.g., Sung, Dhomecha, & Mukhi, 2019)
- large-scale assessments neglected, 2 exceptions (Yamamoto, He, Shin, & von Davier, 2018; Zehner, Sälzer, & Goldhammer, 2016)
## Text Responses with Automatic Processing

**Natural Language Processing and Machine Learning allow to . . .**

- process and classify responses
- access unstructured information for secondary analysis (at the large scale)
- extract information beyond scores

  *(e.g., He, 2013; Zehner, Goldhammer, & Sälzer, 2018)*

**Historically, . . .**

- the start: automatic **essay** grading in 1960s  
  *(Page, 1966, 1968)*
- automatic **short answer** grading started in 1990s  
- **countless developments** since 2010s, two competitions  
  *(Hewlett Foundation, 2012; Dzikovska et al., 2013)*
  *(out-of-date overview at Burrows et al., 2014)*
- **deep learning** and—since 2017—**transformers** in particular  
  *(international)*  
  *(e.g., Sung, Dhomecha, & Mukhi, 2019)*
- **large-scale assessments neglected**, 2 exceptions  
  *(Yamamoto, He, Shin, & von Davier, 2018; Zehner, Sälzer, & Goldhammer, 2016)*
Automatic Scoring of Short Text Responses

In automatic scoring, recent developments take big steps
(e.g., Gong & Yao, 2019; Sung, Dhomecha, & Mukhi, 2019)
Feasibility Requirements in International LSAs
for Automatic Scoring of Short Text Responses

in automatic scoring, recent developments take big steps
(e.g., Gong & Yao, 2019; Sung, Dhomecha, & Mukhi, 2019)
Feasibility Requirements in International LSAs for Automatic Scoring of Short Text Responses

in automatic scoring, recent developments take big steps
(e.g., Gong & Yao, 2019; Sung, Dhamecha, & Mukhi, 2019)

Needs to be Adaptable to . . .

• a vast number of test languages (>50 in PISA)
Feasibility Requirements in International LSAs for Automatic Scoring of Short Text Responses

in automatic scoring, recent developments take big steps (e.g., Gong & Yao, 2019; Sung, Dhamecha, & Mukhi, 2019)

Needs to be Adaptable to...

- a vast number of test languages (>50 in PISA)
- a vast number of items, corresponding coding guides, and multiple domains
Feasibility Requirements in International LSAs for Automatic Scoring of Short Text Responses

in automatic scoring, recent developments take big steps (e.g., Gong & Yao, 2019; Sung, Dhamecha, & Mukhi, 2019)

Needs to be Adaptable to...

- a vast number of test languages (>50 in PISA)
- a vast number of items, corresponding coding guides, and multiple domains
- a vast number of informal and poorly formed texts (low stakes)
Feasibility Requirements in International LSAs
for Automatic Scoring of Short Text Responses

in automatic scoring, recent developments take big steps
(e.g., Gong & Yao, 2019; Sung, Dhamecha, & Mukhi, 2019)

Needs to be Adaptable to . . .

- a vast number of test languages (>50 in PISA)
- a vast number of items, corresponding coding guides, and multiple domains
- a vast number of informal and poorly formed texts (low stakes)
- continuous changes in the assessment design
Feasibility Requirements in International LSAs
for Automatic Scoring of Short Text Responses

in automatic scoring, recent developments take big steps
(e.g., Gong & Yao, 2019; Sung, Dhamecha, & Mukhi, 2019)

Needs to be Adaptable to . . .

- a vast number of test languages (>50 in PISA)
- a vast number of items, corresponding coding guides, and multiple domains
- a vast number of informal and poorly formed texts (low stakes)
- continuous changes in the assessment design

At the Same Time

high-quality coding absolute requirement (i.e., accuracy)
Normalization for Advancing Coding Consistency and Efficiency in PISA

Thus: Let's Unite Two Dissimilar Siblings

<table>
<thead>
<tr>
<th>ReCo</th>
<th>Automatic Text Response Coder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Zehner, Sälzer, &amp; Goldhammer, 2016)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PISA’s MSCS</th>
<th>Machine-Supported Coding System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Yamamoto, He, Shin, &amp; von Davier, 2018)</td>
</tr>
</tbody>
</table>
Thus: Let’s Unite Two Dissimilar Siblings

<table>
<thead>
<tr>
<th>ReCo</th>
<th>Automatic Text Response Coder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Zehner, Sälzer, &amp; Goldhammer, 2016)</td>
</tr>
<tr>
<td>• varying accuracy</td>
<td>(medium to high)</td>
</tr>
<tr>
<td>• perfect coverage of responses</td>
<td></td>
</tr>
<tr>
<td>• as fuzzy as it gets</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PISA’s MSCS</th>
<th>Machine-Supported Coding System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Yamamoto, He, Shin, &amp; von Davier, 2018)</td>
</tr>
</tbody>
</table>
Thus: Let’s Unite Two Dissimilar Siblings

ReCo Automatic Text ResponseCoder
(Zehner, Sälzer, & Goldhammer, 2016)
- varying accuracy (medium to high)
- perfect coverage of responses
- as fuzzy as it gets

PISA’s MSCS Machine-Supported Coding System
(Yamamoto, He, Shin, & von Davier, 2018)
- theoretically, perfect accuracy
- constrained coverage of responses
  (for items with, at least, medium response diversity)
- as picky as it gets
Thus: Let’s Unite Two Dissimilar Siblings

...and Introduce Fuzziness to PISA’s MSCS

ReCo Automatic Text Response Coder

(Zehner, Sälzer, & Goldhammer, 2016)

- varying accuracy (medium to high)
- perfect coverage of responses
- as fuzzy as it gets

PISA’s MSCS Machine-Supported Coding System

(Yamamoto, He, Shin, & von Davier, 2018)

- theoretically, perfect accuracy
- constrained coverage of responses
  (for items with, at least, medium response diversity)
- as picky as it gets
Thus: Let’s Unite Two Dissimilar Siblings
...and Introduce Fuzziness to PISA’s MSCS

<table>
<thead>
<tr>
<th>ReCo</th>
<th>Automatic Text Response Coder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Zehner, Sälzer, &amp; Goldhammer, 2016)</td>
</tr>
<tr>
<td></td>
<td>• varying accuracy (medium to high)</td>
</tr>
<tr>
<td></td>
<td>• perfect coverage of responses</td>
</tr>
<tr>
<td></td>
<td>• as fuzzy as it gets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PISA’s MSCS</th>
<th>Machine-Supported Coding System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Yamamoto, He, Shin, &amp; von Davier, 2018)</td>
</tr>
<tr>
<td></td>
<td>• theoretically, perfect accuracy</td>
</tr>
<tr>
<td></td>
<td>• constrained coverage of responses</td>
</tr>
<tr>
<td></td>
<td>(for items with, at least, medium response diversity)</td>
</tr>
<tr>
<td></td>
<td>• as picky as it gets</td>
</tr>
</tbody>
</table>

Both... take advantage of simple phenomena & baseline methods

thus, easily scalable
PISA’s Machine-Supported Coding System (Yamamoto, He, Shin, & von Davier, 2018)

Starting with PISA 2015’s CBA

- automatically assign code to responses coded before
- pool of coded unique responses (CUR)
  - \( n_{CUR, i} \geq 5 \)
PISA’s Machine-Supported Coding System (Yamamoto, He, Shin, & von Davier, 2018)

Starting with PISA 2015’s CBA

- automatically assign code to responses coded before
- pool of coded unique responses (CUR)
  - $n_{CUR} \geq 5$
  - exact matching, i.e.,
    - each character equal
    - incl. punctuation
    - case-sensitive
PISA’s Machine-Supported Coding System  
(Yamamoto, He, Shin, & von Davier, 2018)

Starting with PISA 2015’s CBA

- automatically assign code to responses coded before
- pool of coded unique responses (CUR)
  - \( n_{CUR} \geq 5 \)
  - exact matching, i.e.,
    - each character equal
    - incl. punctuation
    - case-sensitive

<table>
<thead>
<tr>
<th>Response</th>
<th>Frequencies (full credit)</th>
<th>Frequencies (no credit)</th>
<th>Frequencies (missing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>1,467</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30 minutes</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30mins</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>...</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>( \ldots )</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

High-Level Regularities

- low language diversity
- e.g., 97% coding effort reduction
PISA’s Machine-Supported Coding System (Yamamoto, He, Shin, & von Davier, 2018)

Starting with PISA 2015’s CBA

- automatically assign code to responses coded before
- pool of coded unique responses (CUR)
  - \( n_{CUR} \geq 5 \)
  - exact matching, i.e.,
    - each character equal
    - incl. punctuation
    - case-sensitive

<table>
<thead>
<tr>
<th>Response</th>
<th>Frequencies (full credit)</th>
<th>Frequencies (no credit)</th>
<th>Frequencies (missing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earth Road WF</td>
<td>529</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>earth road WF</td>
<td>76</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>earth road wf</td>
<td>45</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ABC Space Free</td>
<td>0</td>
<td>123</td>
<td>0</td>
</tr>
<tr>
<td>ABC’s Space Free</td>
<td>0</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>ABC’s space free</td>
<td>0</td>
<td>16</td>
<td>0</td>
</tr>
</tbody>
</table>

Medium-Level Regularities
- medium language diversity
- e.g., 61% coding effort reduction

High-Level Regularities
- low language diversity
- e.g., 97% coding effort reduction

(Yamamoto et al., 2018, p. 154)
PISA’s Machine-Supported Coding System (Yamamoto, He, Shin, & von Davier, 2018)

Starting with PISA 2015’s CBA

• automatically assign code to responses coded before
• pool of coded unique responses (CUR)
  o $n_{CUR} \geq 5$
  o exact matching, i.e.,
    • each character equal
    • incl. punctuation
    • case-sensitive

<table>
<thead>
<tr>
<th>Response</th>
<th>Frequencies (full credit)</th>
<th>Frequencies (no credit)</th>
<th>Frequencies (missing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>It states what the paper is going to be about.</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>it tells you what the paper is about</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>its telling you what the paper is about</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>don give up</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>I'dk</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>?</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Low-Level Regularities

• high language diversity
• e.g., 0.4% coding effort reduction

Medium-Level Regularities

• medium language diversity
• e.g., 61% coding effort reduction

High-Level Regularities

• low language diversity
• e.g., 97% coding effort reduction

(Yamamoto et al., 2018, p. 156)
Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response ...

A girl falling into and wandering through a fantasy world.
Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response ...

A girl falling into and wandering through a fantasy world /

... to a numerical representation of its semantics ... (LSA; Deerwester et al., 1990)

\[
\begin{bmatrix}
-0.03 & 0.04 & 0.21 \\
-1.12 & -2.30 & -2.00 & -1.00 \\
0.06 & -0.73 & -0.10 \\
-1.16 & -0.02 & -0.81 \\
-3.37 & 0.04 & -0.51
\end{bmatrix}
\]
Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response ...

\[
\begin{array}{c}
\text{a} \quad \text{girl} \quad \text{falling} \quad \text{into} \quad \text{and} \quad \text{wandering} \quad \text{through} \quad \text{a} \quad \text{fantasy} \quad \text{world} \\
\end{array}
\]
Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response ...

[a] [girl] [falling] [into] [and] [wandering] [through] [a] [fantasy] [world]/
Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response ...

[a] [girl] [falling] [into] [and] [wandering] [through] [a] [fantasy] [world]/
Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response ...

[a] [girl] [falling] [into] [and] [wandering] [through] [a] [fantasy] [world]/
Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response ...

\([a\] [girl] [fall/\ing] [into] [and] [wander/\ing] [through] [a] [fantasy] [world]/\}
Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response ...

[ ] [girl] [fall/ing] [into] and [wander/ing] through [a] [fantasy] [world].

... to a numerical representation of its semantics ... (LSA; Deerwester et al., 1990)

```
<table>
<thead>
<tr>
<th></th>
<th>girl</th>
<th>fall</th>
<th>wander</th>
<th>fantasy</th>
<th>world</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td></td>
<td>−.09</td>
<td>−.11</td>
<td>−.73</td>
<td>−.16</td>
<td>−.07</td>
</tr>
<tr>
<td></td>
<td>.04</td>
<td>.23</td>
<td>.00</td>
<td>.02</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td>.21</td>
<td>.00</td>
<td>−.10</td>
<td>.81</td>
<td>−.51</td>
</tr>
<tr>
<td></td>
<td>−.13</td>
<td>−.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>−.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response ...

... to a numerical representation of its semantics ... (LSA; Deerwester et al., 1990)

... to the automatic code
ReCo vs. MSCS At a Glance
ReCo vs. MSCS At a Glance

Similarities

- group similar responses
ReCo vs. MSCS At a Glance

**Similarities**
- group similar responses
- well-scalable to many languages

**Differences (MSCS vs. ReCo)**
- character vs. semantic level
- no vs. strong normalizing
- perfect vs. varying accuracy
- poor vs. perfect coverage
- poorly vs. easily generalizable across conditions
ReCo vs. MSCS At a Glance

**Similarities**
- group similar responses
- well-scalable to many languages
- build on repeated measurements ("training" data)
ReCo vs. MSCS At a Glance

Similarities
- group similar responses
- well-scalable to many languages
- build on repeated measurements ("training" data)
- item-level
ReCo vs. MSCS At a Glance

**Similarities**
- group similar responses
- well-scalable to many languages
- build on repeated measurements ("training" data)
- item-level

**Differences (MSCS vs. ReCo)**
- character- vs. semantic level
# ReCo vs. MSCS At a Glance

## Similarities
- group similar responses
- well-scalable to many languages
- build on repeated measurements ("training" data)
- item-level

## Differences (MSCS vs. ReCo)
- character- vs. semantic level
- no vs. strong normalizing
- perfect vs. varying accuracy
- poorly vs. easily generalizable across conditions
ReCo vs. MSCS At a Glance

**Similarities**
- group similar responses
- well-scalable to many languages
- build on repeated measurements ("training" data)
- item-level

**Differences (MSCS vs. ReCo)**
- character- vs. semantic level
- no vs. strong normalizing
- perfect vs. varying accuracy
ReCo vs. MSCS At a Glance

Similarities
- group similar responses
- well-scalable to many languages
- build on repeated measurements ("training" data)
- item-level

Differences (MSCS vs. ReCo)
- character- vs. semantic level
- no vs. strong normalizing
- perfect vs. varying accuracy
- poor vs. perfect coverage
ReCo vs. MSCS At a Glance

<table>
<thead>
<tr>
<th>Similarities</th>
<th>Differences (MSCS vs. ReCo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• group similar responses</td>
<td>• character- vs. semantic level</td>
</tr>
<tr>
<td>• well-scalable to many languages</td>
<td>• no vs. strong normalizing</td>
</tr>
<tr>
<td>• build on repeated measurements (“training” data)</td>
<td>• perfect vs. varying accuracy</td>
</tr>
<tr>
<td>• item-level</td>
<td>• poor vs. perfect coverage</td>
</tr>
<tr>
<td></td>
<td>• poorly vs. easily generalizable across conditions</td>
</tr>
</tbody>
</table>
Research Questions

Research Question 1
How does liberating the similarity operationalization affect the automatic coding’s accuracy and reduction of manual coding?
Research Questions

Research Question 1
How does liberating the similarity operationalization affect the automatic coding’s accuracy and reduction of manual coding?

Research Question 2
How generalizable is this across countries/languages?
Data

International Data Complete

- all countries from PISA 2015 and 2018
- 22.6 million text responses in 51 languages from 74 countries
- 233 items from 5 domains
Data

International Data Complete
- all countries from PISA 2015 and 2018
- 22.6 million text responses in 51 languages from 74 countries
- 233 items from 5 domains

Reported Subset
- 85 constructed-response reading items ($n = 2.5$ mio. responses)
- 14 country-by-language groups:
  - English: Australia, Canada, United States
  - French: Canada, France
  - German: Austria, Germany, Italy, Luxembourg, Switzerland
  - Italian: Italy
  - Russian: Russia
  - Spanish: Spain, Chile
Analysis

Subsequent Normalizing Steps

1. Exact Matching
Analysis

<table>
<thead>
<tr>
<th>Subsequent Normalizing Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
</tr>
<tr>
<td>II</td>
</tr>
</tbody>
</table>
## Analysis

### Subsequent Normalizing Steps

<table>
<thead>
<tr>
<th>I</th>
<th>Exact Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>White Space Trimming</td>
</tr>
<tr>
<td>III</td>
<td>Punctuation Removal</td>
</tr>
</tbody>
</table>
### Analysis

#### Subsequent Normalizing Steps

1. Exact Matching
2. White Space Trimming
3. Punctuation Removal
4. Case Insenitivity
5. Spelling Correction
6. Stop Word Removal
7. Low Edit Distance Grouping
8. Synonym Replacement
9. Stemming
10. Bag of Words (i.e., word order neglecting)
11. Semantic Clustering
## Analysis

### Subsequent Normalizing Steps

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Exact Matching</td>
</tr>
<tr>
<td>II</td>
<td>White Space Trimming</td>
</tr>
<tr>
<td>III</td>
<td>Punctuation Removal</td>
</tr>
<tr>
<td>IV</td>
<td>Case Insensitivity</td>
</tr>
<tr>
<td>V</td>
<td>Spelling Correction</td>
</tr>
<tr>
<td>VI</td>
<td>Stop Word Removal</td>
</tr>
</tbody>
</table>
## Analysis

### Subsequent Normalizing Steps

<table>
<thead>
<tr>
<th>I</th>
<th>Exact Matching</th>
<th>VII</th>
<th>Low Edit Distance Grouping</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>White Space Trimming</td>
<td>VIII</td>
<td>Synonym Replacement</td>
</tr>
<tr>
<td>III</td>
<td>Punctuation Removal</td>
<td>IX</td>
<td>Stemming</td>
</tr>
<tr>
<td>IV</td>
<td>Case Insensitivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>Spelling Correction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>Stop Word Removal</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis

Subsequent Normalizing Steps

I Exact Matching
II White Space Trimming
III Punctuation Removal
IV Case Insensitivity
V Spelling Correction
VI Stop Word Removal
VII Low Edit Distance Grouping
VIII Synonym Replacement
IX Stemming
X Bag of Words (i.e., word order neglecting)
XI Semantic Clustering
International Aggregates

Arithmetic Mean

Accuracy
human–computer agreement in %

Human Coding Consistency
% of responses assigned to groups with consistent human coding

Efficiency
% of automatically codable responses

Country–Wise Comparison Across Normalization Steps

Accuracy (M = −0.5%)
Human Coding Consistency (M = −3.9%)
Efficiency (M = +5.1%)
Country & Language A

Accuracy
human–computer agreement in %

Human Coding Consistency
% of responses assigned to groups with consistent human coding

Efficiency
% of automatically codable responses

Semi-Transparent Lines
items

Opaque Lines
arithmetic mean across items
Country & Language B

**Accuracy**
human–computer agreement in %

**Human Coding Consistency**
% of responses assigned to groups with consistent human coding

**Efficiency**
% of automatically codable responses

Semi-Transparent Lines
items

Opaque Lines
arithmetic mean across items
Discussion

**MSCS aka Exact-Matching**

- **substantial savings:** 29% on average
Discussion

MSCS aka Exact-Matching

- substantial savings: 29% on average
- large country-wise differences
Discussion

**MSCS aka Exact-Matching**
- substantial savings: 29% on average
- large country-wise differences

**Improvement By Normalizing**
- substantial gain in efficiency: +5.1%
Discussion

**MSCS aka Exact-Matching**
- **substantial savings**: 29% on average
- **large country-wise differences**

**Improvement By Normalizing**
- **substantial gain in efficiency**: +5.1%
- **minor loss in accuracy**: -0.5%
Discussion

**MSCS aka Exact-Matching**
- **substantial savings**: 29% on average
- **large country-wise differences**

**Improvement By Normalizing**
- **substantial gain in efficiency**: +5.1%
- **minor loss in accuracy**: -0.5%
- **cross-lingual equivalence**: parallel curves promising
**Discussion**

**MSCS aka Exact-Matching**
- **substantial savings**: 29% on average
- **large country-wise differences**

**Improvement By Normalizing**
- **substantial gain in efficiency**: +5.1%
- **minor loss in accuracy**: -0.5%
- **cross-lingual equivalence**: parallel curves promising
- **helpful indicator for human coding consistency**
Cross-Lingual Scoring in International Large-Scale Assessments
Automatic Coding

- Use massive and diverse training data: responses from many languages to build a classifier.
- Do transfer learning: building classifiers for test languages with little data.
- Check human coding consistency across test languages and countries.
- Investigate substantive differences across test languages and countries.
Context

Multi-Lingual Automatic Coding

What if we could...

• use massive and diverse training data: responses from many languages to build a classifier
• do transfer learning: building classifiers for test languages with little data
• check human coding consistency across test languages and countries
• investigate substantive differences across test languages and countries

Nov 3, 2022 | Fabian Zehner | 2022 MARC Conference | Cross-Lingual Scoring in International Large-Scale Assessments
Cross-Lingual Coding: Way More Than Multi-Lingual Automatic Coding
Cross-Lingual Coding: Way More Than Multi-Lingual Automatic Coding

What if we could ...

- use massive and diverse training data: responses from many languages to build a classifier
Cross-Lingual Coding: Way More Than Multi-Lingual Automatic Coding

What if we could ...

- use massive and diverse training data: responses from many languages to build a classifier
- do transfer learning: building classifiers for test languages with little data
Cross-Lingual Coding: Way More Than Multi-Lingual Automatic Coding

What if we could ...

- use **massive and diverse training data**: responses from many languages to build a classifier
- do **transfer learning**: building classifiers for test languages with little data
- check **human coding consistency** across test languages and countries
Cross-Lingual Coding: Way More Than Multi-Lingual Automatic Coding

What if we could ...

- use massive and diverse training data: responses from many languages to build a classifier
- do transfer learning: building classifiers for test languages with little data
- check human coding consistency across test languages and countries
- investigate substantive differences across test languages and countries
The Crux: Capturing Relevant Information and Its Representation

What Makes a Text Response Correct?
(more or less, language-agnostic)

- expressing relevant ...  
  - semantic concepts
  - or proper names
  - rarely, certain character/number sequences

- i.e., propositions, relationships
The Crux: Capturing Relevant Information and Its Representation

What Makes a Text Response Correct? (more or less, language-agnostic)

- expressing relevant...
  - semantic concepts
  - or proper names
  - rarely, certain character/number sequences
- i.e., propositions, relationships

Representing Different Languages: Semantics as the Pivot

- correct
- incorrect
Methodological Approaches

- joint modelling with supervised signal

For Semantic Modelling (see Ruder, Volić, & Søgaard, 2019)
Methodological Approaches

For Semantic Modelling (see Ruder, Vulić, & Søgaard, 2019)

- joint modelling with supervised signal
- joint modelling without supervised signal

Challenges

- cross-lingual and -cultural equivalence
- monitoring quality and potential bias
- constrained semantic spaces in the context of item's topic and focus
- isomorphism assumption, hubness (Ormazabal, Artetxe, Labaka, Soroa, & Agirre, 2019)
Methodological Approaches

For Semantic Modelling (see Ruder, Volić, & Søgaard, 2019)

- joint modelling with supervised signal
- joint modelling without supervised signal
- separate modelling with mapping
Methodological Approaches

For Semantic Modelling (see Ruder, Volić, & Søgaard, 2019)

- **joint modelling** with supervised signal
- **joint modelling** without supervised signal
- **separate modelling** with mapping
- **pre-trained embeddings**; e.g., via transformers (i.e., XLM-R, mBERT)
Methodological Approaches

For Semantic Modelling (see Ruder, Vulić, & Søgaard, 2019)

- joint modelling with supervised signal
- joint modelling without supervised signal
- separate modelling with mapping
- pre-trained embeddings; e.g., via transformers (i.a., XLM-R, mBERT)

Challenges

- cross-lingual and -cultural equivalence
Methodological Approaches

For Semantic Modelling (see Ruder, Volić, & Søgaard, 2019)

- joint modelling with supervised signal
- joint modelling without supervised signal
- separate modelling with mapping
- pre-trained embeddings; e.g., via transformers (i.a., XLM-R, mBERT)

Challenges

- cross-lingual and -cultural equivalence
- monitoring quality and potential bias
Methodological Approaches

For Semantic Modelling (see Ruder, Vulić, & Søgaard, 2019)

- joint modelling with supervised signal
- joint modelling without supervised signal
- separate modelling with mapping
- pre-trained embeddings; e.g., via transformers (i.a., XLM-R, mBERT)

Challenges

- cross-lingual and -cultural equivalence
- monitoring quality and potential bias
- constrained semantic spaces in the context of item’s topic and focus
Methodological Approaches

For Semantic Modelling (see Ruder, Volić, & Søgaard, 2019)
- joint modelling with supervised signal
- joint modelling without supervised signal
- separate modelling with mapping
- pre-trained embeddings; e.g., via transformers (i.a., XLM-R, mBERT)

Challenges
- cross-lingual and -cultural equivalence
- monitoring quality and potential bias
- constrained semantic spaces in the context of item’s topic and focus
- isomorphism assumption, hubness
  (Ormazabal, Artetxe, Labaka, Soroa, & Agirre, 2019)
Supposed Benefit I: Massive and Diverse Training Data

(Zehner, Sälzer, & Goldhammer, 2016, p. 297)

Generalizable Classifiers . . .

- require diverse training data
- diversity comes with data volume
Supposed Benefit I: Massive and Diverse Training Data

(Zehner, Sälzer, & Goldhammer, 2016, p. 297)

Generalizable Classifiers . . .

- require diverse training data
- diversity comes with data volume

Linguistic Variance in Automatic Coding

(Horbach & Zesch, 2019)

- conceptual variance,
- realization variance, and
- non-conformity variance

(Zesch, Horbach, & Zehner, submitted)
Supposed Benefit I: Massive and Diverse Training Data

(Zehner, Sälzer, & Goldhammer, 2016, p. 297)

Generalizable Classifiers . . .

- require **conceptually diverse** training data
- diversity comes with **data volume**

Linguistic Variance in Automatic Coding

(Horbach & Zesch, 2019)

- **conceptual** variance,
- **realization** variance, and
- **non-conformity** variance

(Zesch, Horbach, & Zehner, submitted)
Supposed Benefit I: Massive and Diverse Training Data

(Zehner, Sälzer, & Goldhammer, 2016, p. 297)

- Generalizable Classifiers...
  - require conceptually diverse training data
  - diversity comes with data volume
  - item’s evoked diversity depends on item focus

- Linguistic Variance in Automatic Coding
  (Horbach & Zesch, 2019)
  - conceptual variance,
  - realization variance, and
  - non-conformity variance
  (Zesch, Horbach, & Zehner, submitted)
Supposed Benefit I: Massive and Diverse Training Data

(Zehner, Sälzer, & Goldhammer, 2016, p. 297)

- Generalizable Classifiers . . .
  - require conceptually diverse training data
  - diversity comes with data volume
  - item’s evoked diversity depends on item focus
  - if test language = realization variance, more test languages $\mapsto$ more conceptual variance

Linguistic Variance in Automatic Coding

(Horbach & Zesch, 2019)
- conceptual variance,
- realization variance, and
- non-conformity variance

(Zesch, Horbach, & Zehner, submitted)
Supposed Benefit II: Generalizability (aka Transfer Learning)

Idea

using pre-trained classifiers from other test languages or assessment cycles
Supposed Benefit II: Generalizability (aka Transfer Learning)

Idea

using pre-trained classifiers from other test languages or assessment cycles

So Far, Limited Reported Evidence

- cross-lingual transfer via translation rather weak
  (Horbach, Stennmanns, & Zesch, 2018)
- transfer across cycles rather robust (Zehner & Goldhammer, in press)
Supposed Benefit III: Checking Human Coding Consistency

**Idea**

monitor humans’ coding consistency within and across test languages

---

**Accuracy:** −0.1%

**Consistent Human Coding:** −0.7%

**Effort Reduction:** +6.4%

---

**Accuracy:** −0.2%

**Consistent Human Coding:** −4.0%

**Effort Reduction:** +7.0%
Supposed Benefit III: Checking Human Coding Consistency
Supposed Benefit IV: Contributing to Substantive Research

Idea

granting access to text responses beyond their codes and compare across test languages
Supposed Benefit IV: Contributing to Substantive Research

Idea

granting access to text responses beyond their codes and compare across test languages

So Far, ...

- diverse applications in the mono-lingual space (e.g., Zehner, Goldhammer, & Sälzer, 2018; He, 2013)
- but none in the cross-cultural and -lingual
Supposed Benefit IV: Contributing to Substantive Research

Idea
granting access to text responses beyond their codes and compare across test languages

So Far, ...
- diverse applications in the mono-lingual space (e.g., Zehner, Goldhammer, & Sälzer, 2018; He, 2013)
- but none in the cross-cultural and -lingual
- e.g., explain overall reading literacy across countries and students based on linguistic response features
Literatur


He, Q. (2013). *Text mining and IRT for psychiatric and psychological assessment* (Dissertation, University of Twente, Twente).


Thank You

for your attention