TBA Centre for Technology Based Assessment

DIPF Leibniz Institute for Research and Information in Education

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

DIPF
Leibniz Institute for Research and Information in Education

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

~~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)

~~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

~~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, 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)

~~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, 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)

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
  (Bejar, 1991; Kaplan, 1991; Braun, Bennett, Frye, & Soloway, 1990; Sebrechts, Bennett, & Rock, 1991; Burstein, Kaplan, Wolff, & Lu, 1996)

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
  (Bejar, 1991; Kaplan, 1991; Braun, Bennett, Frye, & Soloway, 1990; Sebrechts, Bennett, & Rock, 1991; Burstein, Kaplan, Wolff, & Lu, 1996)
- 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 (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
  (Bejar, 1991; Kaplan, 1991; Braun, Bennett, Frye, & Soloway, 1990; Sebrechts, Bennett, & Rock, 1991; Burstein, Kaplan, Wolff, & Lu, 1996)
- 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)
- automatic short answer grading started in 1990s
  (Bejar, 1991; Kaplan, 1991; Braun, Bennett, Frye, & Soloway, 1990; Sebrechts, Bennett, & Rock, 1991; Burstein, Kaplan, Wolff, & Lu, 1996)
- 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)
- large-scale assessments neglected, 2 exceptions (Yamamoto, He, Shin, & von Davier, 2018; Zehner, Sälzer, & Goldhammer, 2016)

(international)

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
  (Bejar, 1991; Kaplan, 1991; Braun, Bennett, Frye, & Soloway, 1990; Sebrechts, Bennett, & Rock, 1991; Burstein, Kaplan, Wolff, & Lu, 1996)
- 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)
- large-scale assessments neglected, 2 exceptions (Yamamoto, He, Shin, & von Davier, 2018; Zehner, Sälzer, & Goldhammer, 2016)

(international)

Automatic Scoring of Short Text Responses

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

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)

DIPF
Leibniz Institute for Research and Information in Education

Normalization for Advancing Coding Consistency and Efficiency in PISA

Fabian Zehner, Hyo Jeong Shin, Emily Kerzabi, Nico Andersen, Frank Goldhammer, & Kentaro Yamamoto (2021). *NCME Annual Meeting 2021*.

Thus: Let's Unite Two Dissimilar Siblings

ReCo Automatic Text Response Coder

(Zehner, Sälzer, & Goldhammer, 2016)

PISA's MSCS Machine-Supported Coding System

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

Thus: Let's Unite Two Dissimilar Siblings

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)

Thus: Let's Unite Two Dissimilar Siblings

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

ReCo Automatic Text Response Coder

(Zehner, Sälzer, & Goldhammer, 2016)

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

$\rightleftharpoons$

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

$\rightleftharpoons$

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

Both . . .

take advantage of simple phenomena & baseline methods
$\longmapsto$ 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)
  - $n_{CUR_i} \geq 5$
  - 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)
  - $n_{CUR_i} \geq 5$
  - exact matching, i.e.,
    - each character equal
    - incl. punctuation
    - case-sensitive

Response	Frequencies (full credit)	Frequencies (no credit)	Frequencies (missing)
30	1,467	0	0
30 minutes	23	0	0
30mins	7	0	0
…			
10	0	6	0
5	0	6	0
12	1^2	3	0

(Yamamoto et al., 2018, p. 152)

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)
  - $n_{CUR_i} \geq 5$
  - exact matching, i.e.,
    - each character equal
    - incl. punctuation
    - case-sensitive

Response	Frequencies (full credit)	Frequencies (no credit)	Frequencies (missing)
Earth Road WF	529	I^j	0
earth road WF	76	0	0
earth road wf	45	0	0
...			
ABC Space Free	0	123	0
ABC's Space Free	0	39	0
ABC's space free	0	16	0

(Yamamoto et al., 2018, p. 154)

Medium-Level Regularities

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

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)
  - $n_{CUR_i} \geq 5$
  - exact matching, i.e.,
    - each character equal
    - incl. punctuation
    - case-sensitive

Response	Frequencies (full credit)	Frequencies (no credit)	Frequencies (missing)
It states what the paper is going to be about.	2	I^l	0
it tells you what the paper is about	2	0	0
its telling you what the paper is about	2	0	0
…			
don give up	0	2	0
Idk	0	2	0
?	0	1	0

(Yamamoto et al., 2018, p. 156)

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

Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response …

```
A girl falling into and wandering through a fentasy world .
```

Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response …

`A girl falling into and wandering through a fentasy world /`

Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response …

a girl falling into and wandering through a fentasy world /

Employed Automatic Coding (Zehner et al., 2016)

Example: Starting with a short text response …

`[a] [girl] [falling] [into] [and] [wandering] [through] [a] [fentasy] [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]/

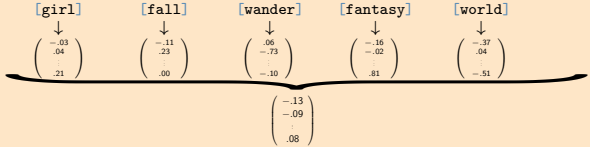
DIPF
Leibniz Institute for Research and Information in Education

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]./

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

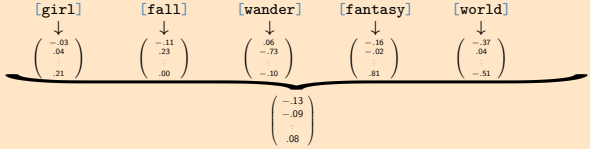


Employed Automatic Coding (Zehner et al., 2016)

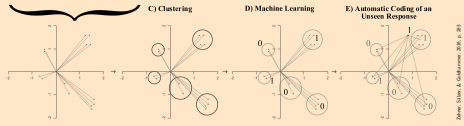
Example: Starting with a short text response …

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

… 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

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

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

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
- poorly vs. easily generalizable across conditions

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

I Exact Matching

Analysis

Subsequent Normalizing Steps

I Exact Matching
II White Space Trimming

Analysis

Subsequent Normalizing Steps

I Exact Matching
II White Space Trimming
III Punctuation Removal

Analysis

Subsequent Normalizing Steps

I Exact Matching
II White Space Trimming
III Punctuation Removal
IV Case Insenitivity

Analysis

Subsequent Normalizing Steps

I Exact Matching
II White Space Trimming
III Punctuation Removal
IV Case Insenitivity
V Spelling Correction
VI Stop Word Removal

Analysis

Subsequent Normalizing Steps

I Exact Matching
II White Space Trimming
III Punctuation Removal
IV Case Insenitivity
V Spelling Correction
VI Stop Word Removal
VII Low Edit Distance Grouping
VIII Synonym Replacement
IX Stemming

Analysis

Subsequent Normalizing Steps

I Exact Matching
II White Space Trimming
III Punctuation Removal
IV Case Insenitivity
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

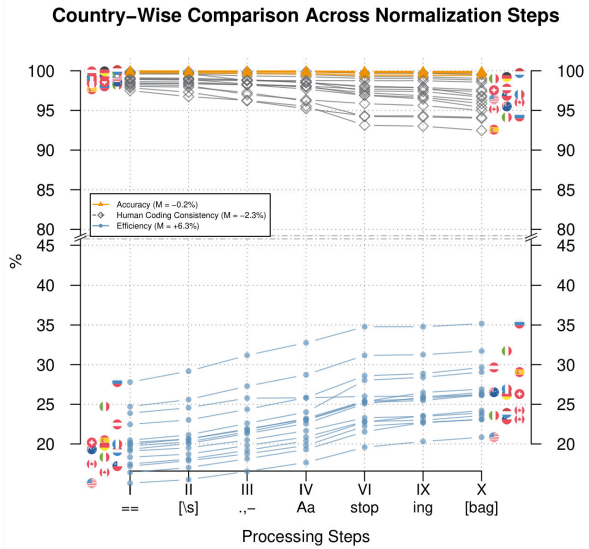
International Aggregates

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 & Language A

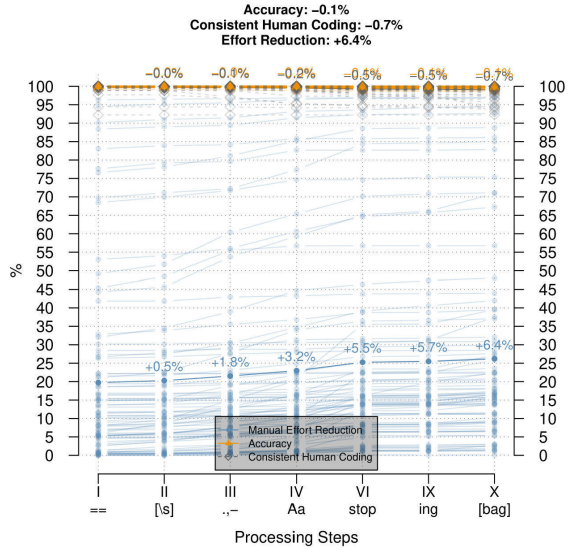
Accuracy
human–computer agreement in %

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

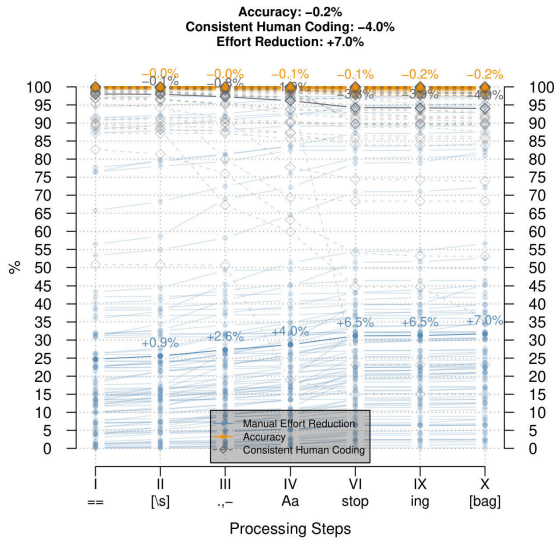
Accuracy
human–computer agreement in %

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

DIPF
Leibniz Institute for Research and Information in Education

Cross-Lingual Scoring
in International Large-Scale Assessments

DIPF
Leibniz Institute for Research and Information in Education

Automatic Coding

DIPF
Leibniz Institute for Research and Information in Education

Multi-Lingual Automatic Coding

DIPF
Leibniz Institute for Research and Information in Education

Cross-Lingual Coding: Way More Than Multi-Lingual Automatic Coding

DIPF
Leibniz Institute for Research and Information in Education

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

DIPF
Leibniz Institute for Research and Information in Education

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

DIPF
Leibniz Institute for Research and Information in Education

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

DIPF
Leibniz Institute for Research and Information in Education

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

DIPF
Leibniz Institute for Research and Information in Education

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

DIPF
Leibniz Institute for Research and Information in Education

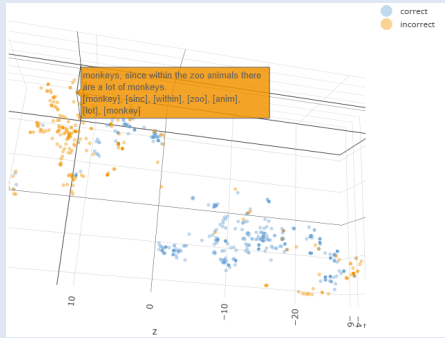
The Crux: Capturing Relevant Information and Its Representation

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



DIPF
Leibniz Institute for Research and Information in Education

Methodological Approaches

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

- joint modelling with supervised signal

DIPF
Leibniz Institute for Research and Information in Education

Methodological Approaches

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

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

DIPF
Leibniz Institute for Research and Information in Education

Methodological Approaches

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

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

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)

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, 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

DIPF
Leibniz Institute for Research and Information in Education

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

DIPF
Leibniz Institute for Research and Information in Education

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)

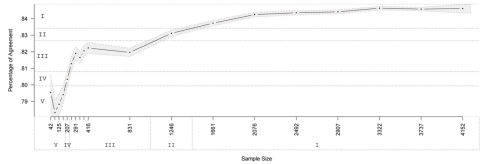
Challenges

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

DIPF
Leibniz Institute for Research and Information in Education

Supposed Benefit I: Massive and Diverse Training Data

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



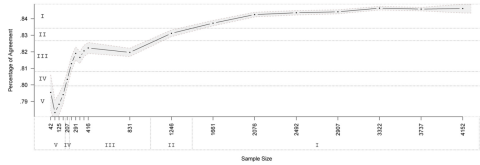
Generalizable Classifiers . . .

- require diverse training data
- diversity comes with data volume

DIPF
Leibniz Institute for Research and Information in Education

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

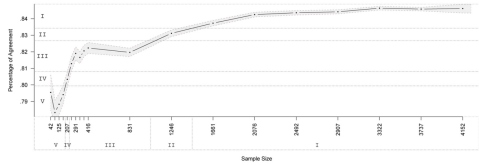
(Horbach & Zesch, 2019)

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

DIPF
Leibniz Institute for Research and Information in Education

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

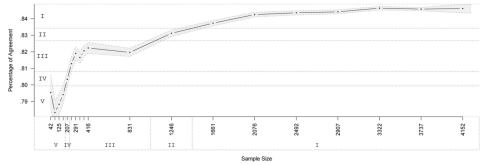
(Horbach & Zesch, 2019)

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

DIPF
Leibniz Institute for Research and Information in Education

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

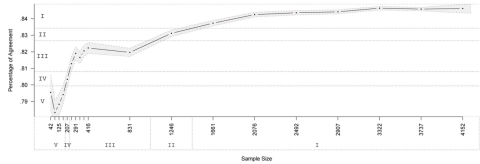
(Horbach & Zesch, 2019)

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

DIPF
Leibniz Institute for Research and Information in Education

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)

DIPF
Leibniz Institute for Research and Information in Education

Supposed Benefit II: Generalizability (aka Transfer Learning)

Idea

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

DIPF
Leibniz Institute for Research and Information in Education

Supposed Benefit II: Generalizability (aka Transfer Learning)

Idea

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

So Far, Limited Reported Evidence

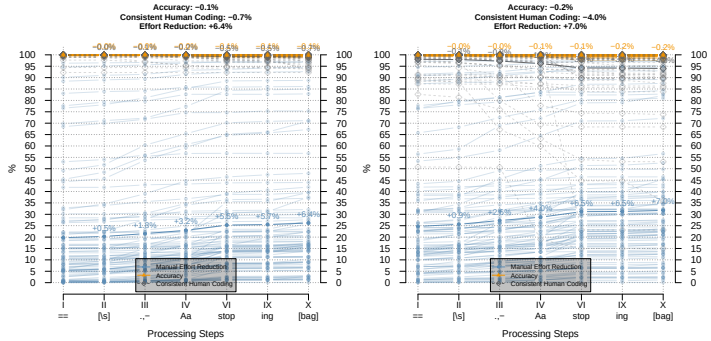
- cross-lingual transfer via translation rather weak
  (Horbach, Stennmanns, & Zesch, 2018)
- transfer across cycles rather robust (Zehner & Goldhammer, in press)

DIPF
Leibniz Institute for Research and Information in Education

Supposed Benefit III: Checking Human Coding Consistency

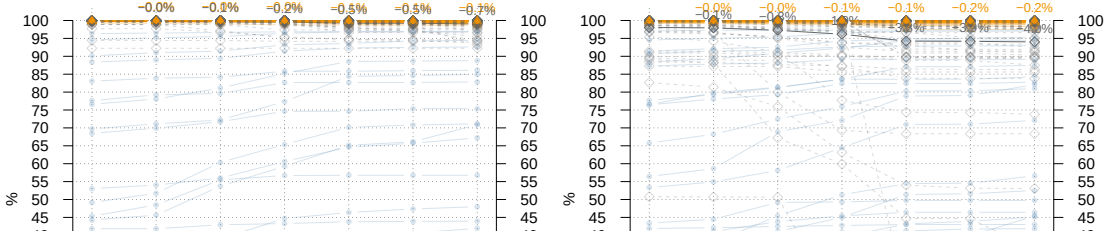
Idea

monitor humans' coding consistency within and across test languages



DIPF
Leibniz Institute for Research and Information in Education

Supposed Benefit III: Checking Human Coding Consistency



DIPF
Leibniz Institute for Research and Information in Education

Supposed Benefit IV: Contributing to Substantive Research

Idea

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

DIPF
Leibniz Institute for Research and Information in Education

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

DIPF
Leibniz Institute for Research and Information in Education

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

DIPF
Leibniz Institute for Research and Information in Education

Literatur

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Gong, T., & Yao, X. (2019). An attention-based deep model for automatic short answer score. *International Journal of Computer Science and Software Engineering*, *8*(6), 127–132.

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

Horbach, A., Stennmanns, S., & Zesch, T. (2018). Cross-lingual content scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 410–419).

Horbach, A., & Zesch, T. (2019). The Influence of Variance in Learner Answers on Automatic Content Scoring. *Frontiers in Education*, *4*, 4.

Ormazabal, A., Artetxe, M., Labaka, G., Soroa, A., & Agirre, E. (2019). *Analyzing the Limitations of Cross-lingual Word Embedding Mappings*.

Ruder, S., Vulić, I., & Søgaard, A. (2019). A Survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, *65*, 569–631. Retrieved from `https://www.jair.org/index.php/jair/article/view/11640`

Sung, C., Dhamecha, T. I., & Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education* (pp. 469–481). Cham: Springer International Publishing.

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2018). Development and implementation of a machine-supported coding system for constructed-response items in PISA. *Psychological Test and Assessment Modeling*, *60*(2), 145–164.

Zehner, F., & Goldhammer, F. (in press). Automatically Coding PISA Text Responses. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative Computer-based International Large-Scale Assessments – Foundations, Methodologies and Quality Assurance Procedures*. Cham: Springer.

Zehner, F., Goldhammer, F., & Sälzer, C. (2018). Automatically analyzing text responses for exploring gender-specific cognitions in PISA reading. *Large-Scale Assessments in Education*, *6:7*.

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280–303.

Zesch, T., Horbach, A., & Zehner, F. (submitted). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses.

DIPF
Leibniz Institute for Research and Information in Education

Thank You

for your attention

TBA Zentrum für technologiebasiertes Assessment

andersen.nico@dipf.de fabian.zehner@dipf.de

@fabian_zehner