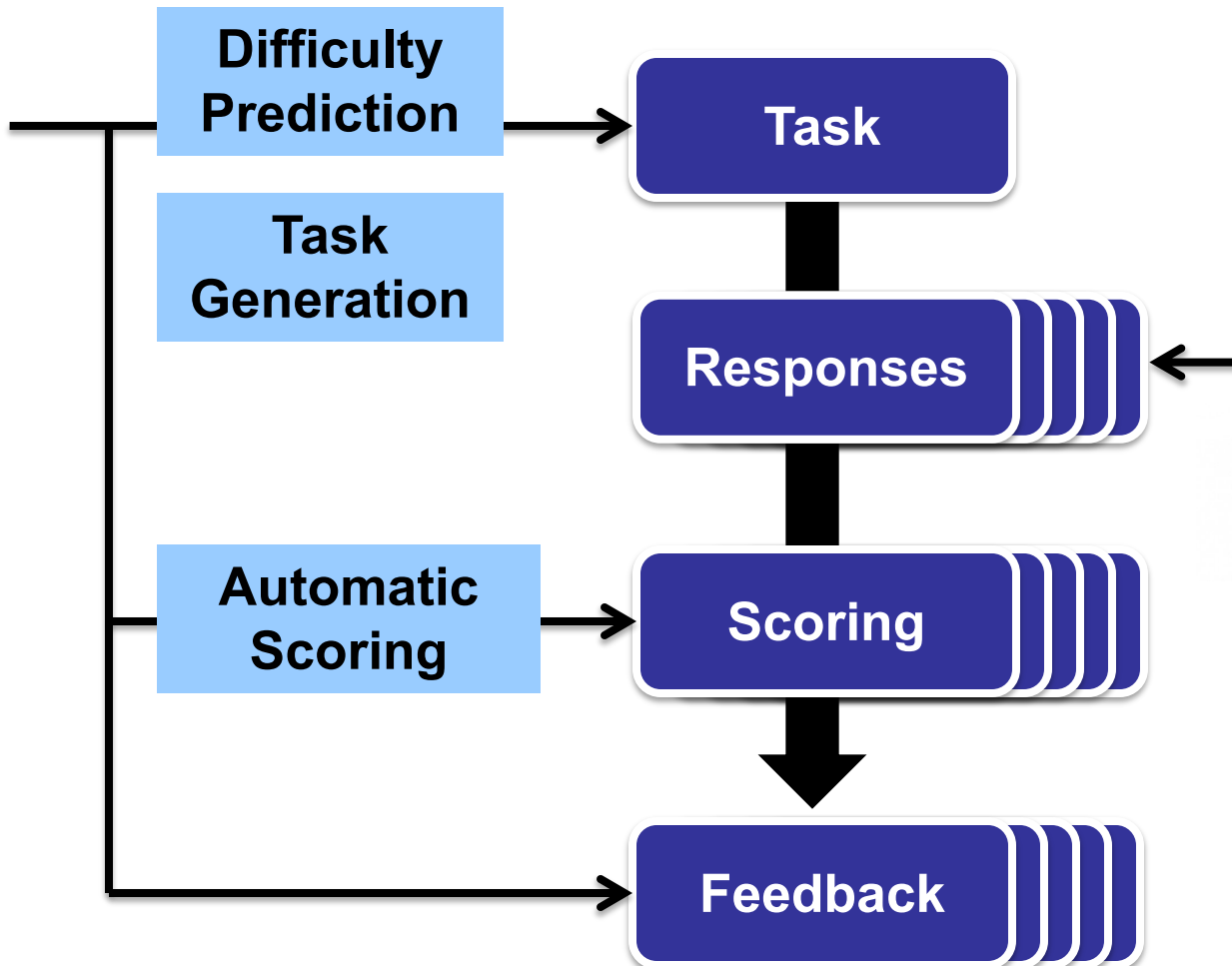


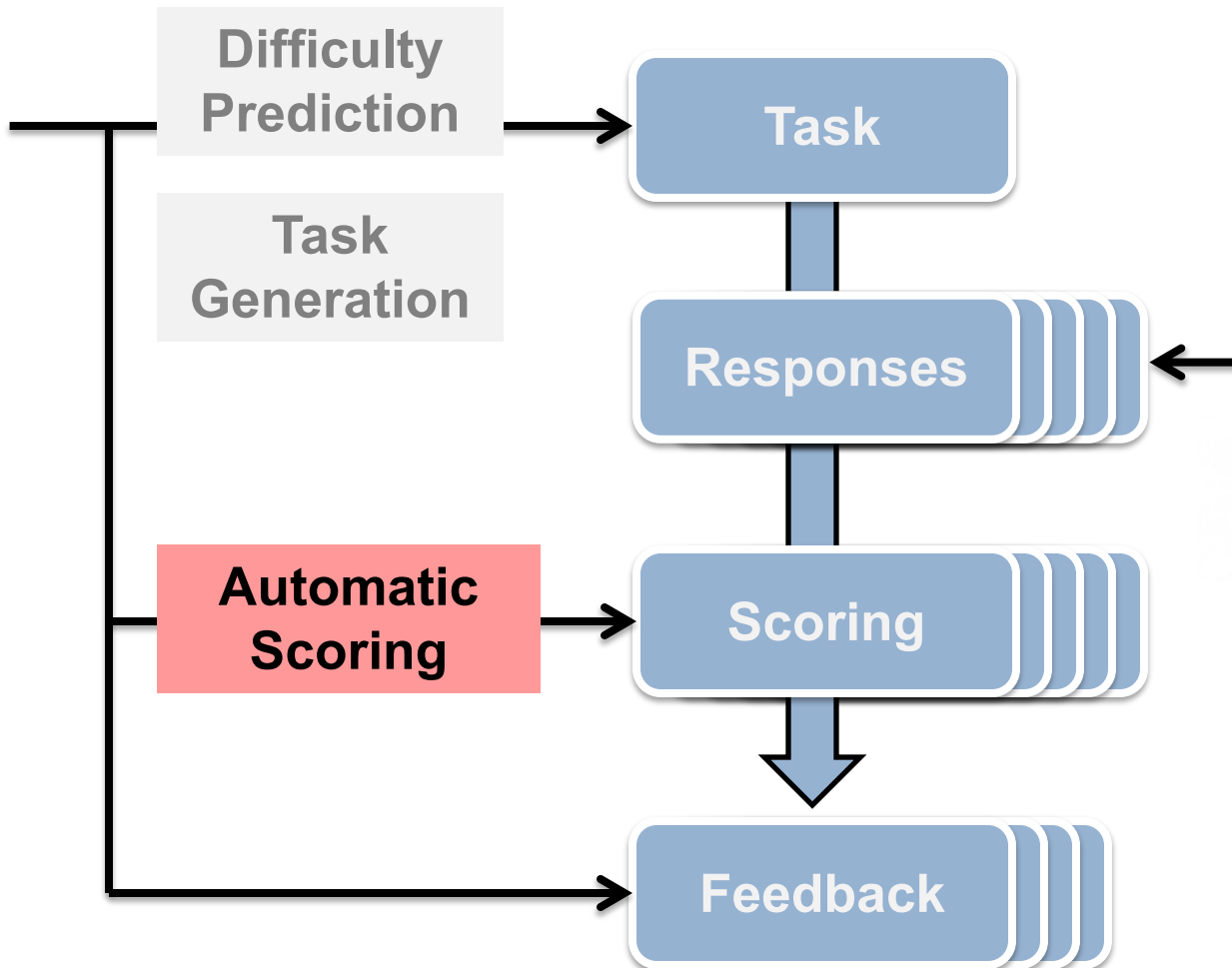
Innovative Tools in der Hochschulbildung – Freitext-Assessment

Marie Bexte
Andrea Horbach
Torsten Zesch

Language Technology Lab
University of Duisburg-Essen

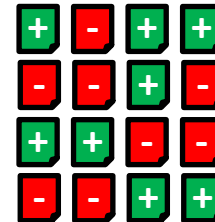
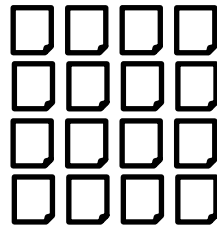




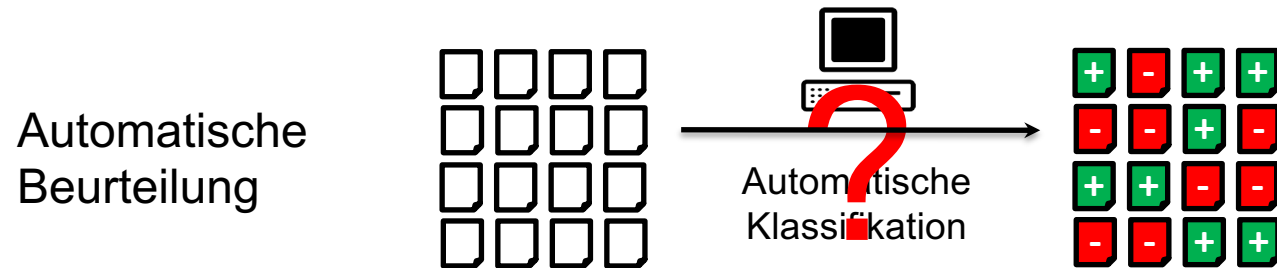


Automatische Beurteilung

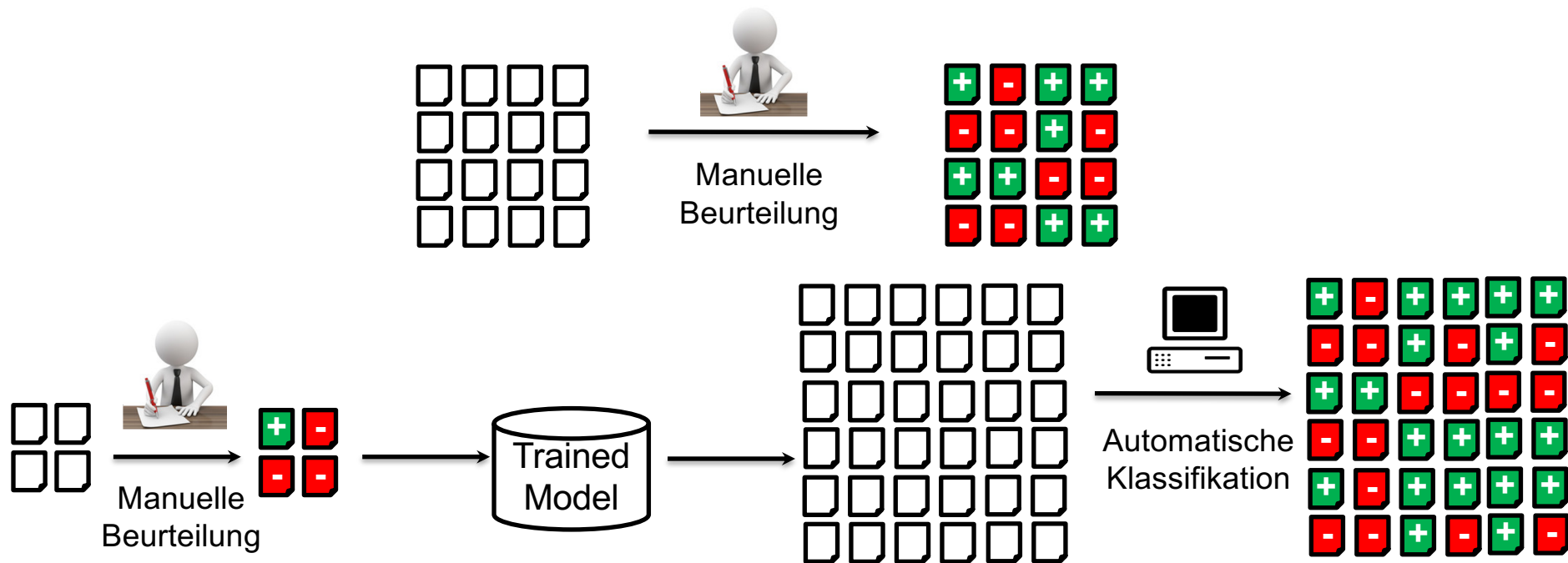
Manuelle
Beurteilung



Automatische Beurteilung



Automatische Beurteilung



Antwortlänge



Frage: *Wo wurde Mozart geboren?*

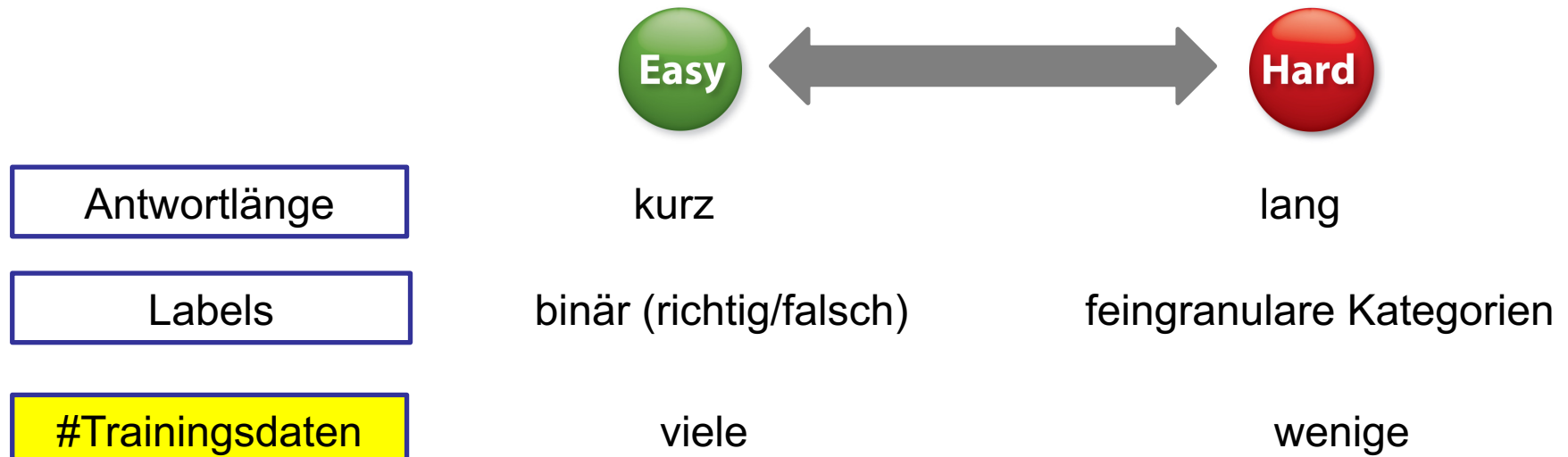
Antworten:

Salzburg.

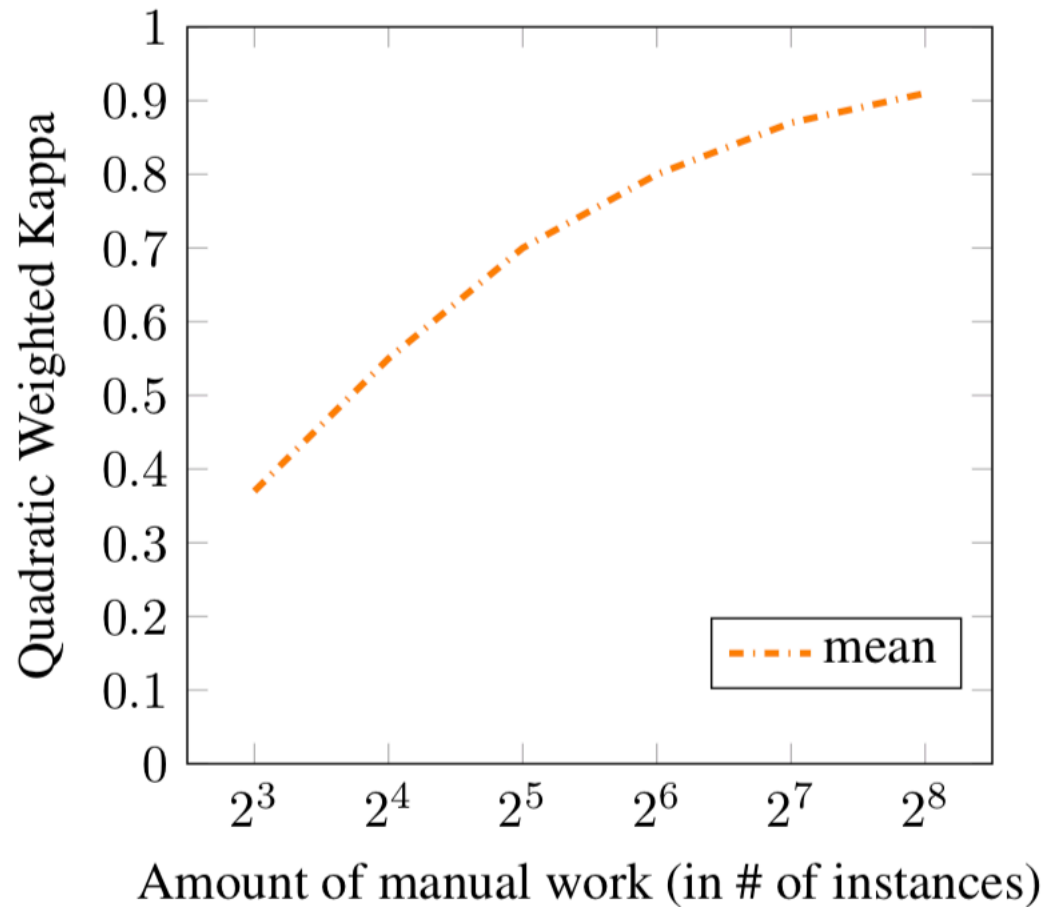
*Mozart wurde in Salzburg geboren.
1756 in Wien.*

Frage: *Wie funktioniert Photosynthese?*

Antwort: *Pflanzen sind zur Photosynthese fähig. Dabei wandeln sie mit Hilfe von Chlorophyll Lichtenergie in chemische Energie um. Aus Kohlendioxid und Wassser wird so (...)*

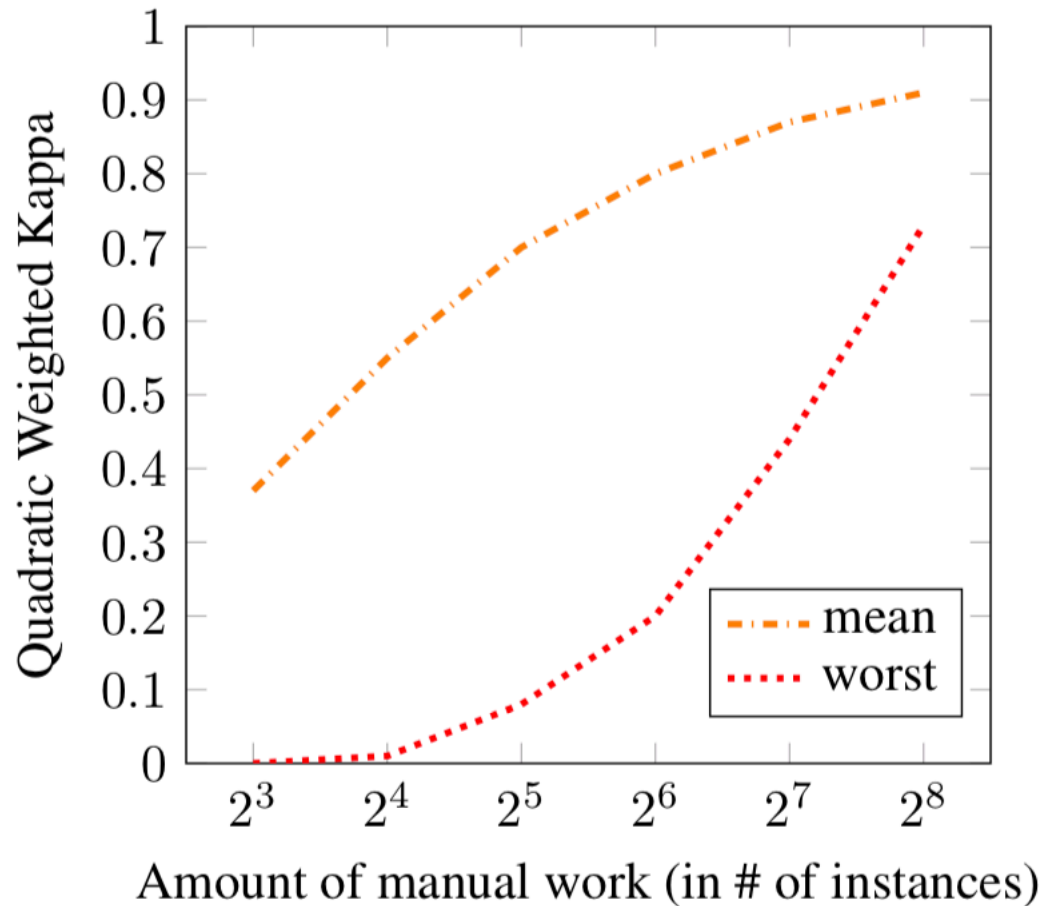


Menge an Trainingsdaten



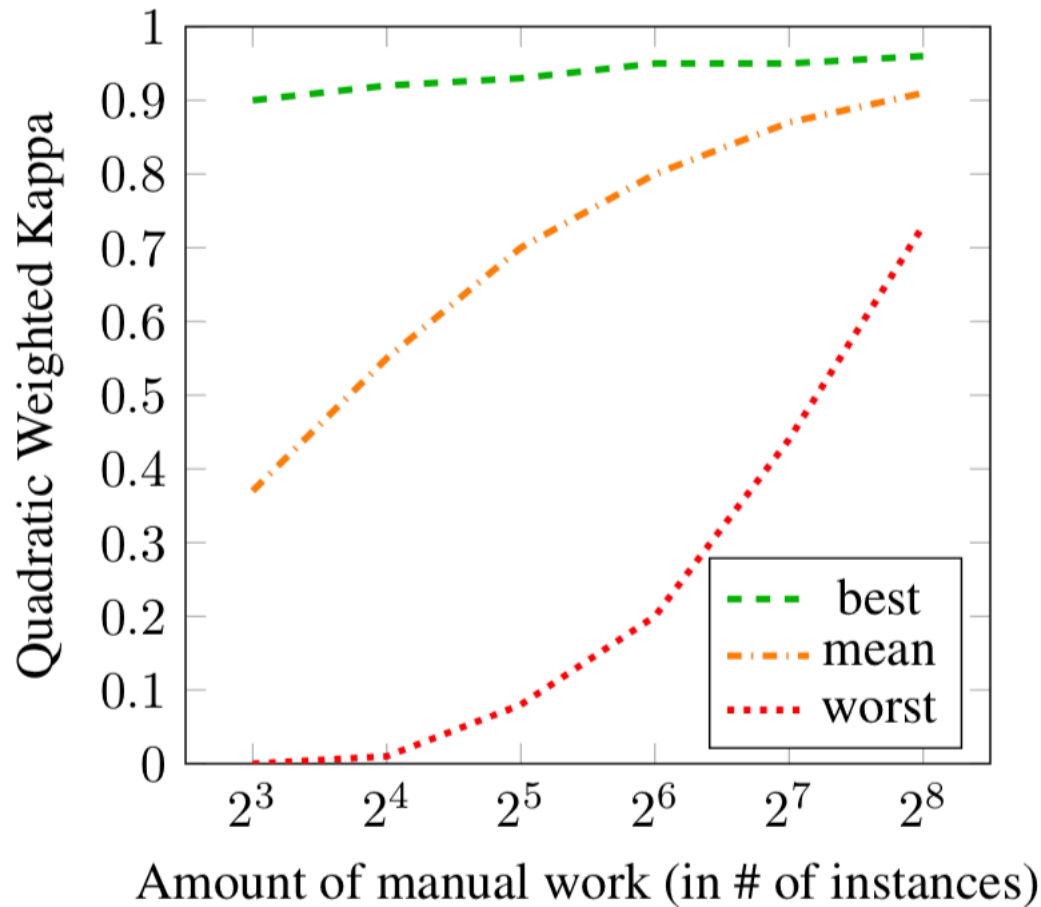
Reducing Annotation Efforts in Supervised Short Answer Scoring. Torsten Zesch, Michael Heilman, Aoife Cahill, BEA@NAACL 2015

Menge an Trainingsdaten



Reducing Annotation Efforts in Supervised Short Answer Scoring. Torsten Zesch, Michael Heilman, Aoife Cahill, BEA@NAACL 2015

Menge an Trainingsdaten



Reducing Annotation Efforts in Supervised Short Answer Scoring. Torsten Zesch, Michael Heilman, Aoife Cahill, BEA@NAACL 2015



Modalität

am Computer
geschrieben

handgeschrieben/
gesprochen

Modalität – Bewertung handschriftlicher Antworten

Original

Panda bears are similar to koala bears because one, they are both a type of bear and two, they eat grass & plants. They are different than pythons because pythons are reptile and they eat things whole.

Gold-Standard Transkription

“Panda bears are similar to koala bears because one, they are both a type of bear and two, they eat grass & plants. They are different than pythons because pythons are reptile and they eat things whole.”

Word Segmentation

Panda bear are similar to koala bears because one, they are both a type of bear and two, they eat grass & plants. They are different than pythons because pythons are reptile and they eat things whole.

Ergebnis der Handschrifterkennung

Paonde bears are fimiter to houla bears hecauge one , they ae bad a tyee of hear and tusr they erat gre l plant ftay are eslptnt them pgton heeause puythorr ove reprite and they eat thrings whale

Modalität – Bewertung handschriftlicher Antworten

Pre-processing	Post-processing	BLEU	QWK
lower bound (random)			-.07
Basic	-	3	.21
Full	-	24	.30
Oracle	-	27	.42
Oracle	GENERIC Dict	37	.40
Oracle	NOISY Dict	36	.45
Oracle	CLEAN Dict	24	.53
upper bound (digitized)		100	.65

handschriftlicher Text
mit unterschiedlichem
Spellchecking

computer-
geschriebener
Text



Modalität

Sprache

am Computer
geschrieben
monolingual

handgeschrieben/
gesprochen
cross-lingual

Cross-lingual Scoring

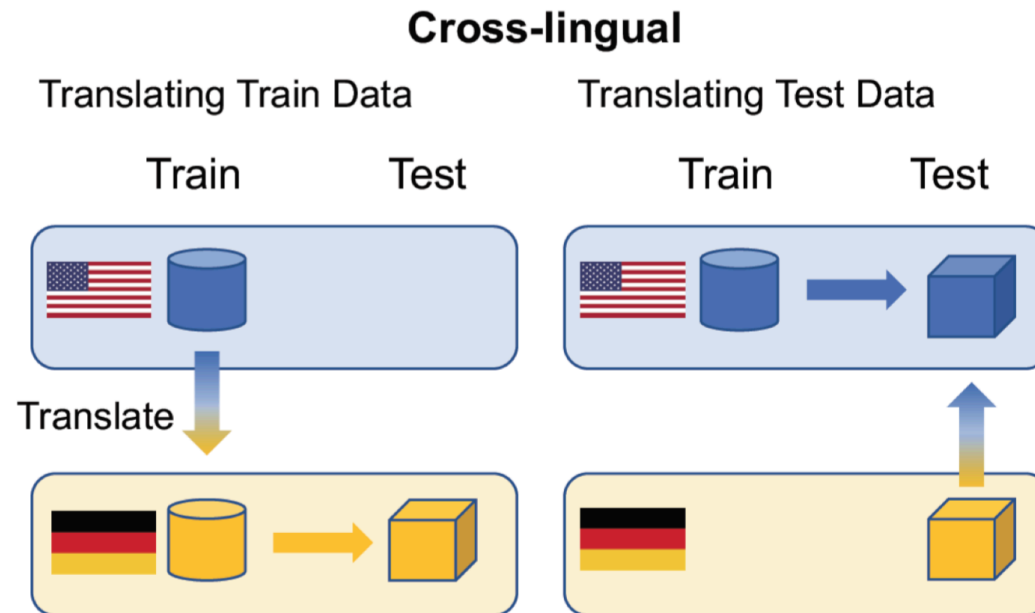
Monolingual

Train

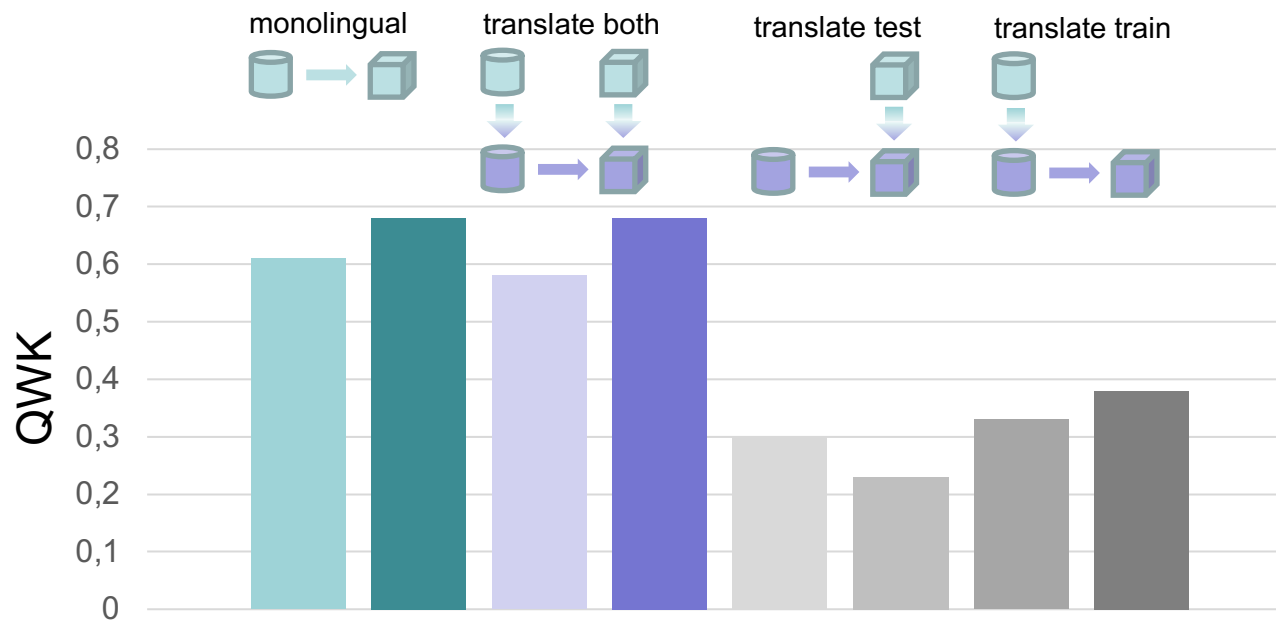
Test



Cross-lingual Scoring



Cross-lingual Scoring



ASAP prompt 1,2,10



Modalität

am Computer
geschrieben

handgeschrieben/
gesprochen

Sprache

monolingual

cross-lingual

Explizitheit

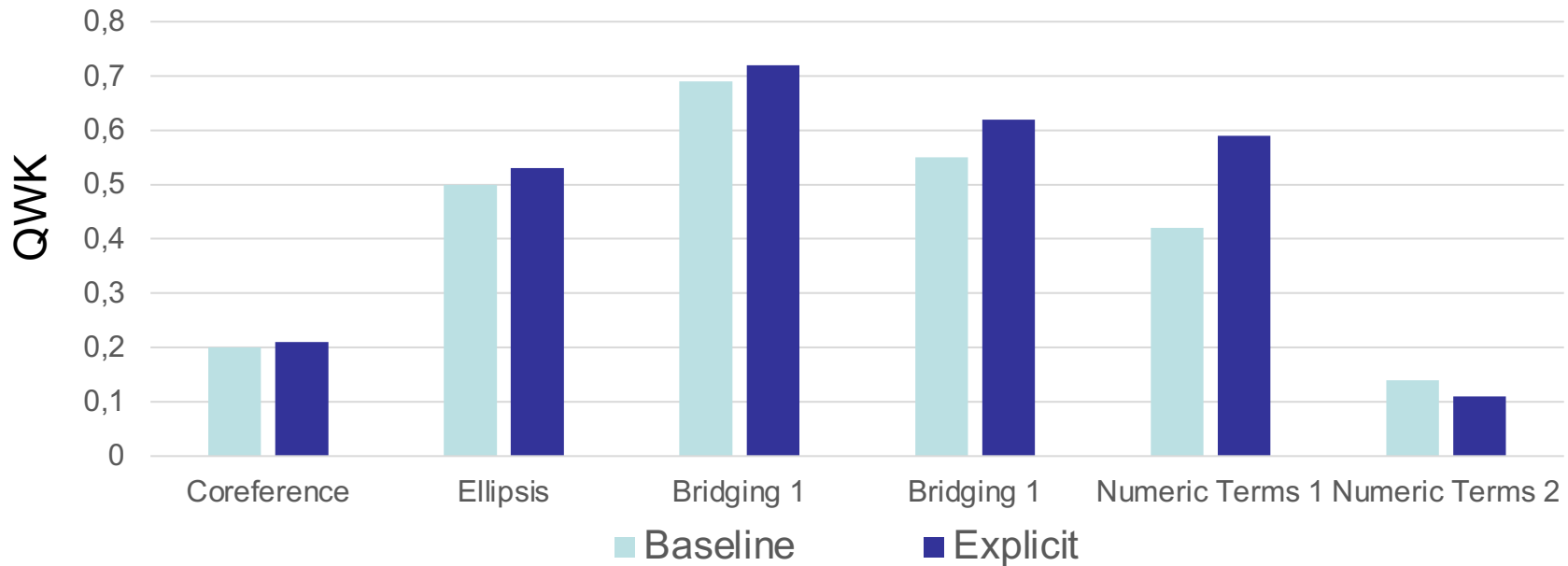
explicit

implicit

Implizite vs explizite Sprache

Implizite Antwort: *Pandas eat bamboo. They can reach an age as old as 30.*

Explizite Antwort: *Pandas eat bamboo. **Pandas** can reach an age as old as 30 years.*



Implicit Phenomena in Short-answer Scoring Data. Marie Bexte, Andrea Horbach, Torsten Zesch. (under review)



Modalität

am Computer
geschrieben

handgeschrieben/
gesprochen

Sprache

monolingual

cross-lingual

Explizitheit

explicit

implicit

Cheating

selten

häufig

Wie anfällig sind Bewertungsmodelle für Adversarials?



Frage: *Explain how pandas in China are similar to koalas in Australia...* (ASAP Prompt 3)

Richtige Antwort: (...) *A panda only eats bamboo, and a Koala bear only eats eucalyptus leaves.*

Verschiedene Arten von Adversarials

- **Random characters:** *fcwowtpbambqal wkjxldrla befhg*
- **Random words:** *a line of then and perhaps one calculation in eating*
- **Random prompt-specific words:** *environment however in Australia because support their diet panda or koala*
- **“Content Burst”:** *panda bamboo koala eucalyptus China Australia*
- **Shuffle:** *bamboo eucalyptus eats koalas need eat nothing only*
- **GPT:** *Pandas and Koalas are similar in their*

Ergebnis:
Bis zu 60% der Adversarials werden vom System akzeptiert.

Don't take "nswvtnvakgxpm" for an answer - The surprising vulnerability of
Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill and Torsten Zesch. COLING



Modalität

am Computer
geschrieben

handgeschrieben/
gesprochen

Sprache

monolingual

cross-lingual

Explizitheit

explicit

implicit

Cheating

selten

häufig

Erklärbarkeit
/Feedback

egal

wichtig

- Ob automatisches Scoring funktioniert, hängt von vielen Faktoren ab
- Meta-Constraints: automatic vs. assisted scoring, high vs low stakes Tests

Easy

Hard

Länge	×	
Labels	×	
Anzahl Trainingsdaten	×	
Modalität		×
Sprache	×	
Explizitheit	×	
Cheating		×
Erklärbarkeit / Feedback	×	

Beispiel: Klausur an der Uni



Publications LTL – Automatic Scoring

- **C-Test Collector: A Proficiency Testing Application to Collect Training Data for C-Tests** Christian Haring, René Lehmann, Andrea Horbach, Torsten Zesch. BEA@EACL2021
- **Appropriateness and Pedagogic Usefulness of Reading Comprehension Questions** Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle and Montse Maritxalar. LREC 2020.
- **Exploring the Impact of Handwriting Recognition on the Automated Scoring of Handwritten Student Answers.** Christian Gold and Torsten Zesch. ICFHR 2020.
- **Don't take "nswvtnvakgxp" for an answer - The surprising vulnerability of automatic content scoring systems to adversarial input.** Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill and Torsten Zesch. COLING 2020.
- **Chinese Content Scoring: Open-Access Datasets and Features on Different Segmentation Levels** Yuning Ding, Andrea Horbach, Haoshi Wang, Xuefeng Song and Torsten Zesch. AACL-IJCNLP 2020.
- **The Influence of Variance in Learner Answers on Automatic Content Scoring.** Andrea Horbach and Torsten Zesch. *Frontiers in Education*, volume 4, 2019.
- **Cross-lingual Content Scoring.** Andrea Horbach, Sebastian Stennmanns, Torsten Zesch. BEA@NAACL 2018
- **ESCRITO - An NLP-Enhanced Educational Scoring Toolkit.** Torsten Zesch and Andrea Horbach. LREC 2018
- **The Influence of Spelling Errors on Content Scoring Performance.** Andrea Horbach, Yuning Ding, Torsten Zesch, NLPTEA 2017
- **Investigating neural architectures for short answer scoring** Brian Riordan, A. Horbach, Aoife Cahill, T. Zesch and Chong Min Lee, BEA@EMNLP, 2017.
- **Fine-grained essay scoring of a complex writing task for native speakers.** A. Horbach, Dirk Scholten-Akoun, Y. Ding and T. Zesch, BEA@EMNLP, 2017.
- **Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks.** Ildikó Pilán, Elena Volodina and Torsten Zesch, COLING 2016
- **Task-Independent Features for Automated Essay Grading,** Torsten Zesch, Michael Wojatzki, Dirk Scholten-Akoun, BEA@NAACL 2015

Publications LTL – Task Generation

- **Reducing Annotation Efforts in Supervised Short Answer Scoring**, Torsten Zesch, Michael Heilman, Aoife Cahill, BEA@NAACL 2015
- **Towards Automatic Scoring of Cloze Items by Selecting Low-Ambiguity Contexts**, T. Horsmann, T. Zesch, NEALT Proceedings Series Vol. 22, 33-42, 2014
- **UKP-BIU: similarity and entailment metrics for student response analysis**, Torsten Zesch, Omer Levy, Iryna Gurevych, Ido Dagan, SemEval-2013
- **A flexible online system for curating reduced redundancy language exercises and tests**. Torsten Zesch, Melanie Goggin, Andrea Horbach, Jennifer Wrede-Jackes: EuroCALL 2018
- **The Role of Diacritics in Designing Lexical Recognition Tests for Arabic**. Osama Hamed and Torsten Zesch, ACLing 2017
- **Validating Bundled Gap Filling – Empirical Evidence for Ambiguity Reduction and Language Proficiency Testing Capabilities** Niklas Meyer, Michael Wojatzki and Torsten Zesch, In Proceedings of the NLP4CAL at SLTC 2016, 2016.
- **Bundled Gap Filling: A New Paradigm for Unambiguous Cloze Exercises**. Michael Wojatzki, Oren Melamud and Torsten Zesch, BEA@NAACL, 2016
- **Predicting the Spelling Difficulty of Words for Language Learners** Lisa Beinborn and Torsten Zesch, BEA Workshop at NAACL, 2016.
- **Generating Nonwords for Vocabulary Proficiency Testing**. Osama Hamed, Torsten Zesch, Language Technology Conference, 2015.
- **Candidate Evaluation Strategies for Improved Difficulty Prediction of Language Tests**. Lisa Beinborn, T. Zesch, Iryna Gurevych, BEA@NAACL 2015
- **Predicting the Difficulty of Language Proficiency Tests** Lisa Beinborn, Torsten Zesch, Iryna Gurevych, Transactions of the ACL, 517-529, 2014
- **Automatic generation of challenging distractors using context-sensitive inference rules**, Torsten Zesch, Oren Melamud, BEA Workshop at ACL 2014