Computational Linguistic Analysis, Assessment, and Language Development

Considering Language and Task

Detmar Meurers
University of Tübingen

http://icall-research.de

EALTA 2019
Dublin

Introduction

- Where could computational linguistic methods be relevant?
  - automating aspects of summative assessment
    - → saving effort/cost, improving consistency of scoring
    - → allowing item types that are more open in large-scale studies
  - formative assessment
    - → immediate, interactive feedback in Tutoring Systems
    - → adaptive materials: selecting or generating items, supporting individual learning paths
  - interpreting large-scale language data
    - → more representative samples
    - → analyzing longitudinal development
  - interpreting authentic language-use data
    - → data arising “in the wild”
    - → learning process data (interaction logs, . . .)

Introduction

- Computational linguistics focuses on analyzing language.
  - But how valid can the interpretation of language be without considering task and people characteristics?
- Language testing has developed sophisticated methods for designing items and interpreting responses.

  But this also is not without challenges:
  - discarding ill-behaved items ≠ understanding language and task properties needed to predict item difficulty
  - psychometric quality ≠ ecological validity
  - how to benefit from observations of authentic language use and process data of learning “in the wild”

⇒ Fostering the cross-disciplinary discussion between the fields could be fruitful in linking different expertises.

Introduction

Some existing synergies

- Automatic essay scoring is widely used in high-stakes testing (e.g., TOEFL, GRE) and in some school systems.
- An increasingly broad range of CL methods is used, with Latent Semantic Analysis playing a central role:
  - Compare a new essay to already graded essays in terms of which words occur in it.
  - Classify an essay in relation to a reference set of essays.
- How about computational linguistic methods directly characterizing proficiency and language development?

  → Linguistic complexity analysis

Linguistic complexity

- Complexity is “a matter of the number and variety of an item’s constituent elements and of the elaborateness of their interrelational structure”. (Rescher 1998)
- Skehan (1989) characterized proficiency in terms of the three dimensions Complexity, Accuracy & Fluency (CAF)
  - Complexity is “the extent to which the language produced in performing a task is **elaborate** and **varied**.” (Ellis 2003)

Evidence about linguistic complexity

I. Which language **forms** are used, how are they combined?
   - type of forms in the **linguistic system**
     e.g. complex NPs per sentence
     → theoretical linguistics
   - use of forms in **individual language experience**
     e.g. word frequency in corpora representative of experience
     → corpus linguistics, psychology

II. What type and amount of **meaning** is encoded by the forms and how is it organized into a coherent **discourse**?
     e.g. propositional idea density, referential cohesion
     → models of reading in psychology

III. What are the demands on **human processing**?
     e.g. memory load, surprisal/expected continuation
     → psycholinguistics

→ CL methods support the identification of such features

(e.g., McNamara et al. 2014; Kyle 2016; Lu 2014; Vajjala 2015; Chen 2018)

Analyzing development in large-scale data

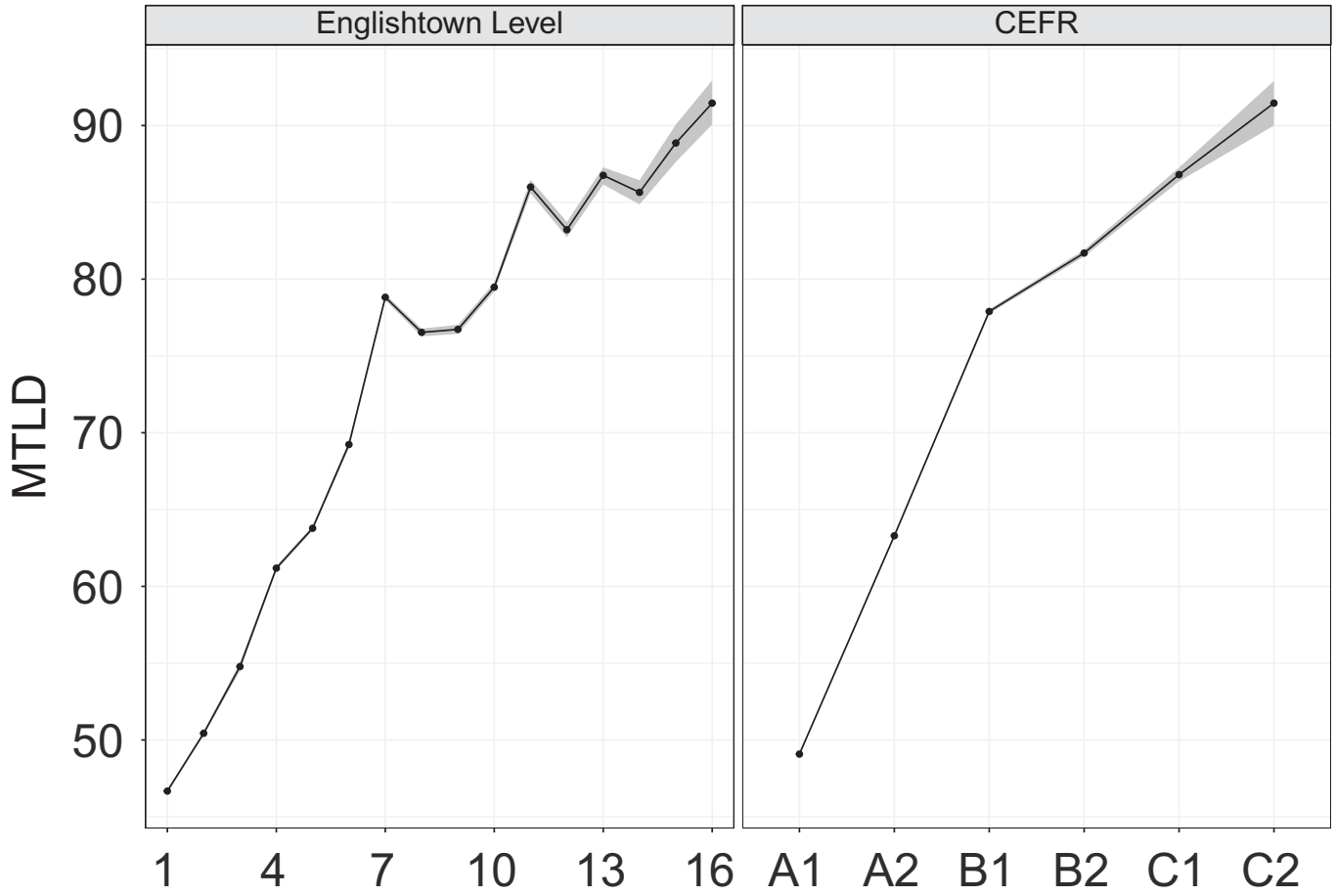
- Computer-based language learning supports the collection of large-scale learner data:
  - English Town system → EFCamDat (Geertzen et al. 2013)
- The EFCamDat corpus (v. 2) contains
  - 175 thousand learners, who provided 1.18 million texts
  - 16 levels covering A1 to C2 in the CEFR

→ Empirically grounded analyses of development based on

- aspects of the language system, e.g., relative clauses (Alexopoulou, Geertzen, Korhonen & Meurers 2015)
- general characterization of development, e.g., complexity (Alexopoulou, Michel, Murakami & Meurers 2017)
  - CL methods used for complexity analysis (Vajjala & Meurers 2012) are available online at http://purl.org/ctap (Chen & Meurers 2016)

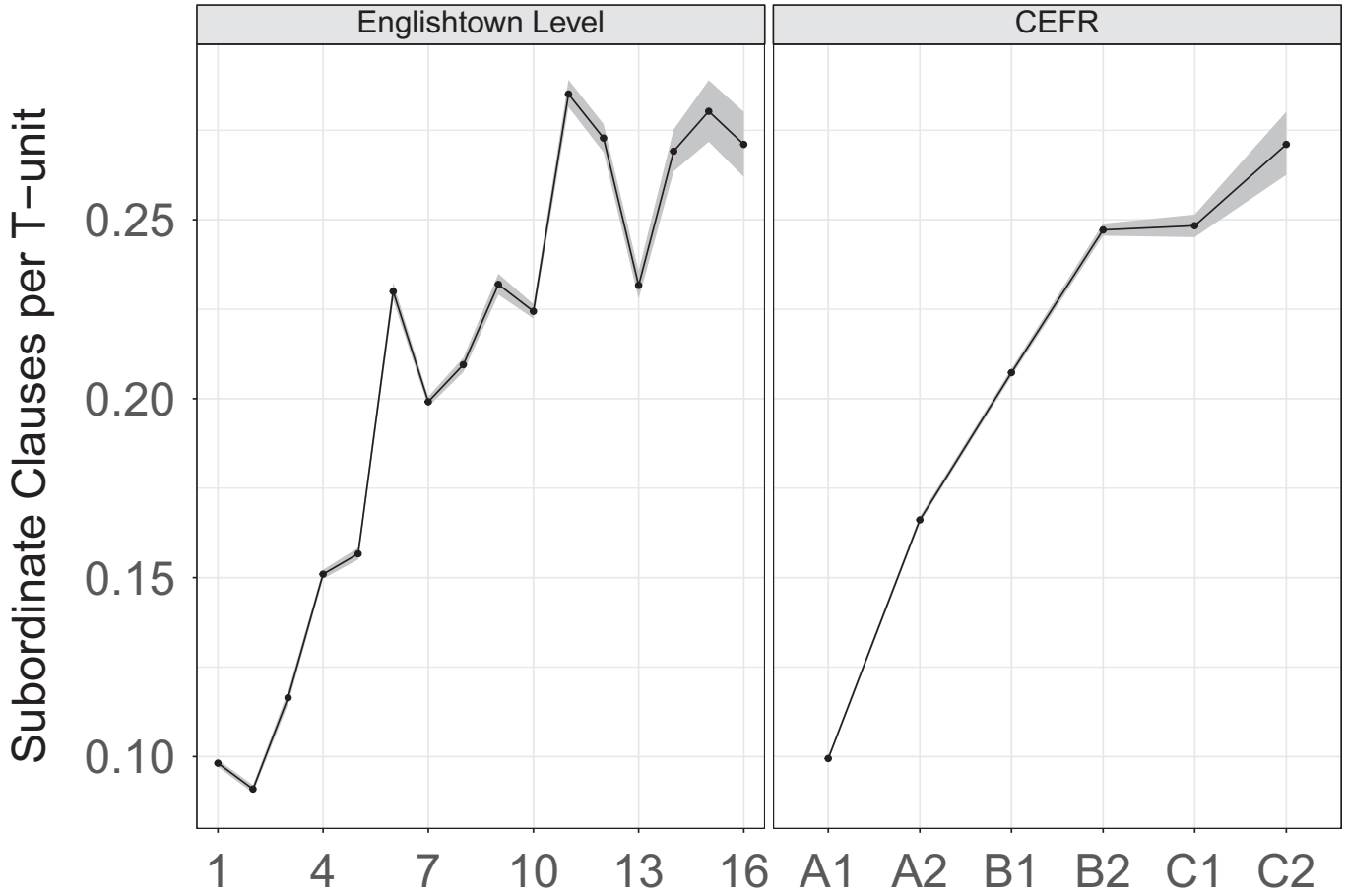
Complexity development in EFCamDat

Lexical Diversity



Complexity development in EFCamDat

Syntactic elaborateness: subordination



Complexity development in EFCamDat

- Large-scale learner corpus analysis using NLP methods confirms global development of linguistic complexity.
- But aren't we missing something?
  - We analyze the **linguistic complexity** of texts that are
  - produced by people differing in **language proficiency**.
- How about the nature of the **tasks** that the language is produced for?

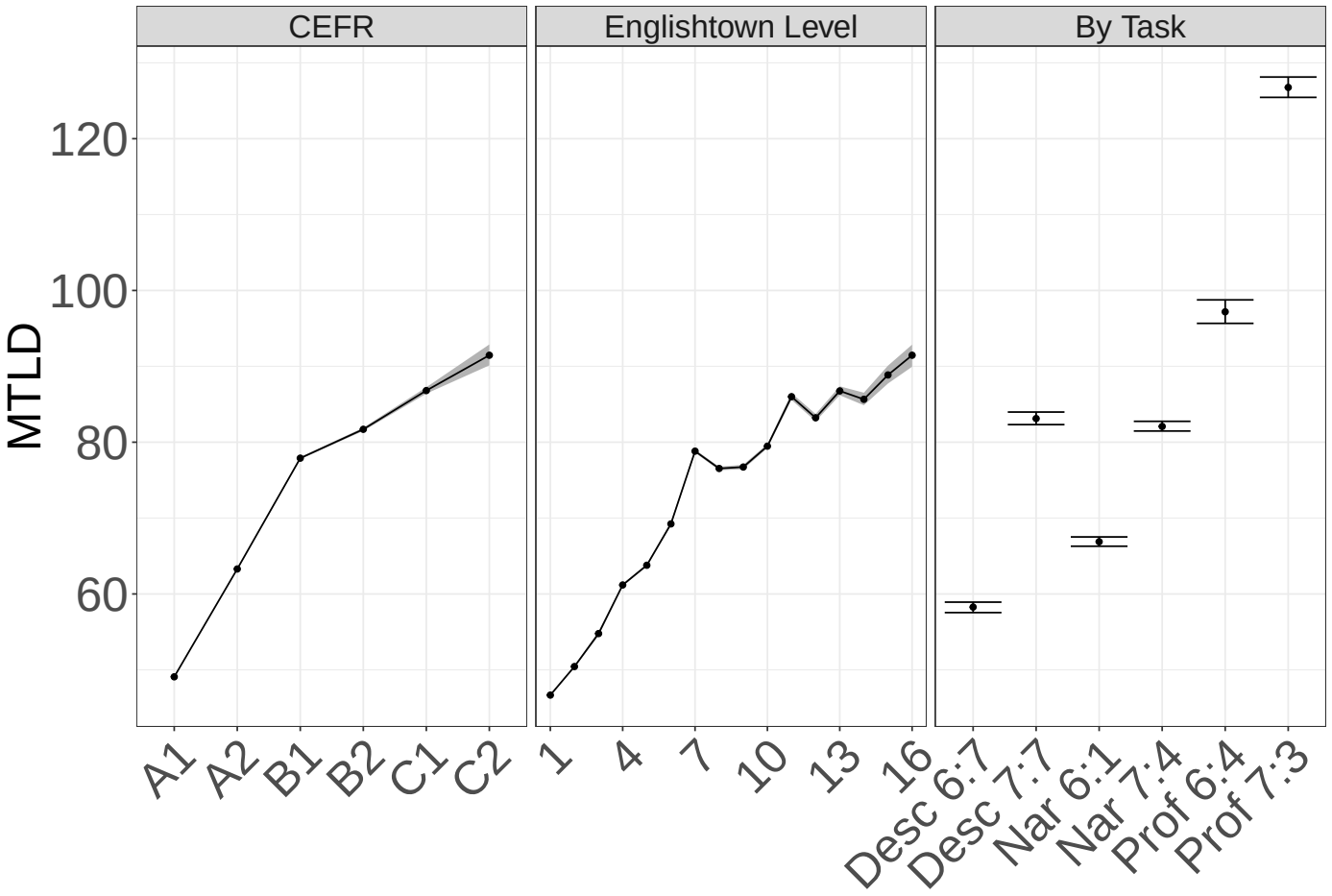
How about the tasks?

(Alexopoulou, Michel, Murakami & Meurers 2017)

- Impact of Task Complexity?
  - Limited Attentional Capacity Model (Skehan 1998)
  - Cognition Hypothesis (Robinson 1995)
- Impact of Task Type?
- We focus on three task types at the intermediate level (B1)
  - **Description**: letter of complaint about a meal, a cruise
  - **Narrative**: write a movie plot, about an experience
  - **Professional**: write a resume, a job advertisement

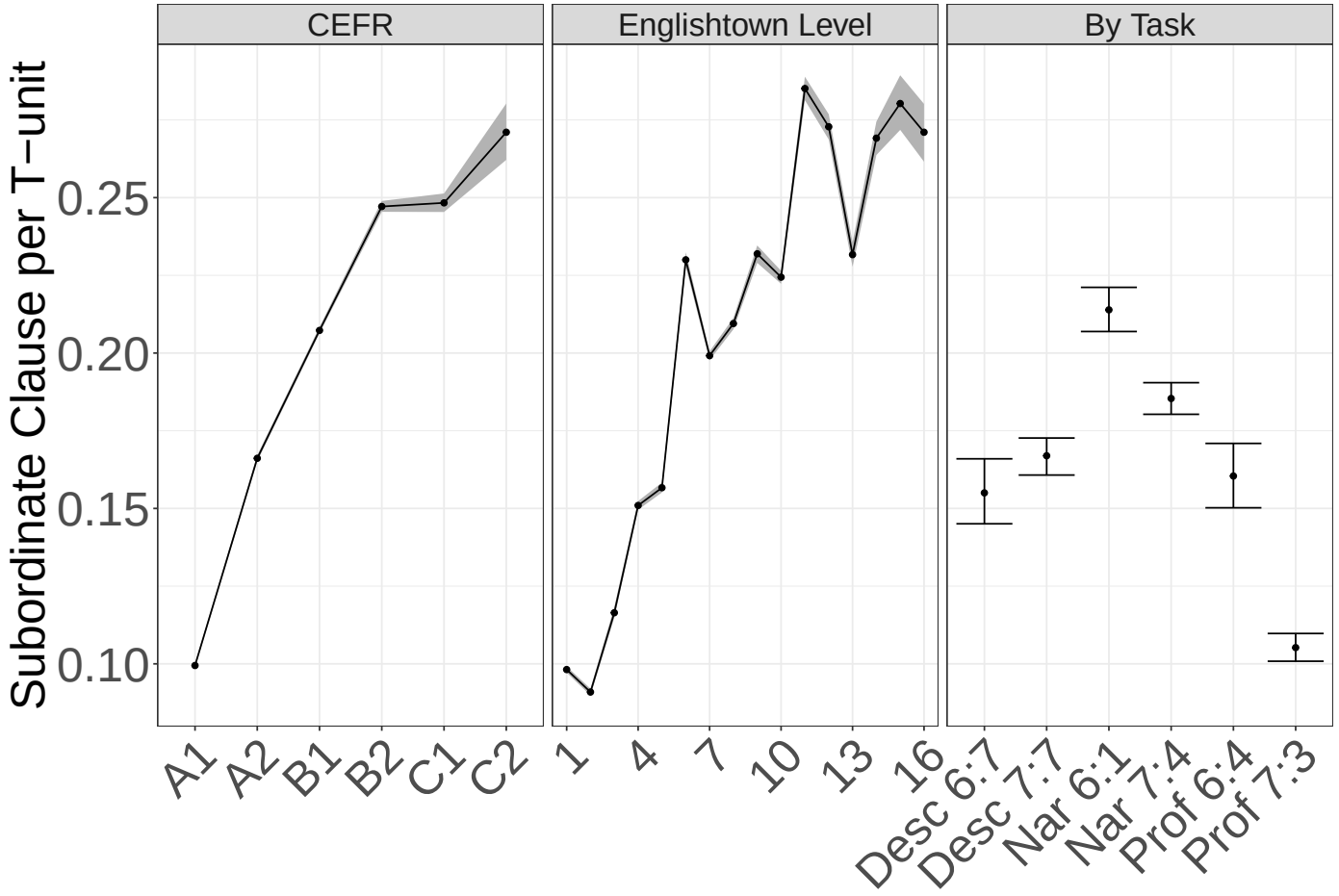
Complexity and task effects in EFCamDat

Lexical Diversity



Complexity and task effects in EFCamDat

Syntactic elaborateness: subordination



Zooming in on tasks

- Tasks can impact complexity more than development.
  - a “particularly severe threat to validity” (Vyatkina 2012)

  → Representativity of tasks in learner corpora (Wisniewski 2017)

  → Study task impact using learner corpora (Karges et al. in press)

→ Tasks need more attention.

  - Model tasks in terms of the nature and variability of language production they elicit. (Quixal & Meurers 2016)
    - Is the variability in line with the learning goals?
    - Can we provide valid interpretations (automatically)?
  - Model difficulty-defining characteristics of items to support generation of exercises with predicted difficulty, e.g.,
    - C-test gap difficulty prediction (Svetashova 2015; Beinborn 2016)
    - Cued gap-fill item difficulty prediction (Pandarova et al. 2019)
  - Understanding language-based task demands also key for other subject domains, e.g., mathematical word problems.

Manipulating task complexity

(Daroczy, Wolska, Meurers & Nuerk 2015)

- Mathematical word problems differ in mathematical task complexity (operation, carry, congruence, . . .):

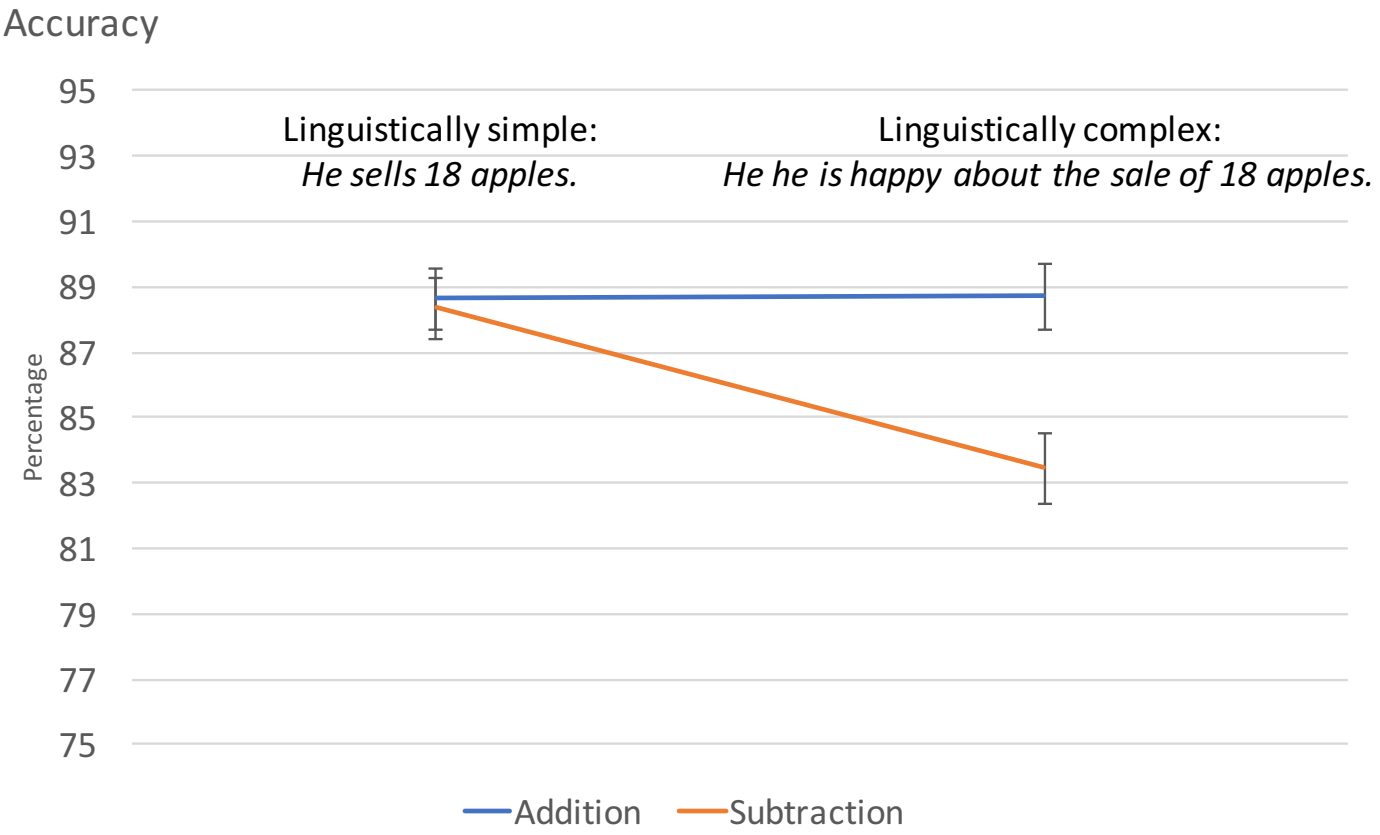
  (1) *A farmer went to market. He arrived with 47 apples.*
  a. *He sells 5 apples.*
  b. *He sells 18 apples.*
  How many apples does he have left?

- What happens if we orthogonally manipulate the linguistic complexity? Here: nominalization

  (2) a. *He sells 18 apples.*
  b. *He is happy about the sale of 18 apples.*

Manipulating task complexity

(Daroczy, Meurers, Heller, Wolska & Nuerk under review)



N=331, school children (mean age=10,91, SD=1,26)

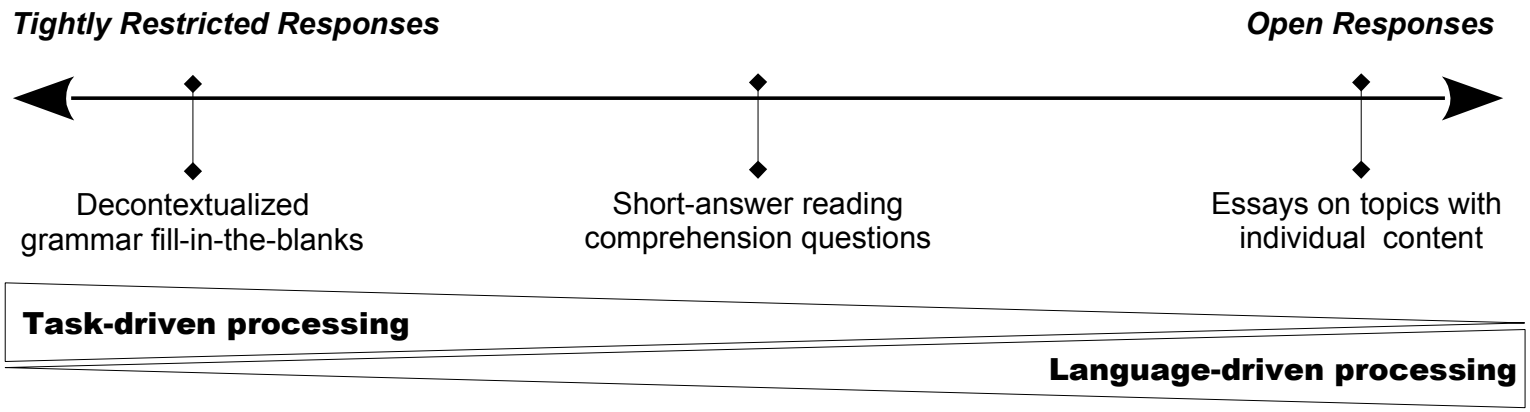
- Language complexity and mathematical task complexity interact, even for conceptually independent parameters.

Interpreting learner language

- Given only the language system, how would you analyze
  - (3) *I don't know where he live.*
  - (4) *I didn't know*
  - (5) *I don't know his lives.*
  - (6) *I know where he lives.*
  - (7) *I know he lived*
- These are from a translation task in the Hiroshima English Learners' Corpus (Miura 1998), for the Japanese of
  - (8) *I don't know where he lives.*

$\Rightarrow$ Valid interpretation requires explicit task / language context.

Exploring task-driven processing



- Let's explore using CL methods to analyze learner responses starting from the tasks:
  - → formative assessment to provide feedback in a tutoring system for English in real-life secondary school

Task-driven processing in an ITS

A web-based workbook for English: FeedBook

- Starting point: printed **Camden Town** Workbook
  - approved for 7th grade English classes in Germany
  - fully integrated into real-life education
  - workbook exercises mostly assigned as homework
- The FeedBook is a web-based implementation of the printed workbook, which in addition
  - provides immediate scaffolded feedback guiding learners.
  - Coverage:
    - all 7th grade grammar topics (form-oriented)
    - reading/listening exercises (meaning-oriented)
- Analysis (Rudzewitz et al. 2017, 2018) starts from the task model,
  - generating possible well-formed and ill-formed variation based on the set of target answers, which then is
  - compared to learner answer using flexible matching.

82 Gillian's point of view

Complete this version of the story from Gillian's point of view. Use the following verbs and put them in the simple past.

- *LiF1Rc: Simple past*

begin · come · feel · get · give · go · lie · make · not be · not listen · put · say · sit · suggest · try

Support
Irregular verbs:
Textbook p. 255

Mum's boyfriend was coming to meet me so of course I *got* ✔ up in a bad mood. But Mum *gave* ✔ me a great big smile. She *made* ✔ me my favourite pancakes with maple syrup for breakfast but I *wasn't* ✔ hungry. She *tryed* to cheer me up and _____ that we go shopping. That usually puts me in a good mood but not today. So I _____ something about homework and _____ into my room. I _____ down on my bed and _____ really sorry for myself. Just then Mum _____ in. She

B2 Gillian's point of view

Complete this version of the story from Gillian's point of view. Use the following verbs and put them in the simple past.

• LiF1Rc: Simple past

begin · come · feel · get · give · go · lie · make · not be · not listen · put · say · sit · suggest · try

Support
Irregular verbs:
Textbook p. 255

Mum's boyfriend was coming to meet me so of course I *got* ✓ up in a bad mood. But Mum *gave* ✓ me a great big smile. She *made* ✓ me my favourite pancakes wit ✓ hungry. She *tryed* ✗ to shopping. That usually puts me in a good mood. Nothing about homework and into my room. I down on my bed and really sorry for myself. Just then Mum in. She

Feedback für "tryed"

When an infinitive ends in 'consonant + y', we change the 'y' to 'i' in the simple past.

Hilfreich? Ja Nein

OK

C 3 What was ... doing while Gillian was doing something else?

Write down what Gillian's friends were doing while she was running away from home. Use the past progressive in both parts of the sentence.

• *LiFTRe: Past progressive*

buy Arsenal tickets | feed Patch | watch TV | sit on the bus

1. buy Arsenal tickets/sit on the bus

Charlie was buying Arsenal tickets while Gillian was sitting on the bus.

2. feed Patch/sit on the bus

George fed Patch while Gillian was sitting on the bus.

3. watch TV/sit on the bus

8.1 Off to Greece again

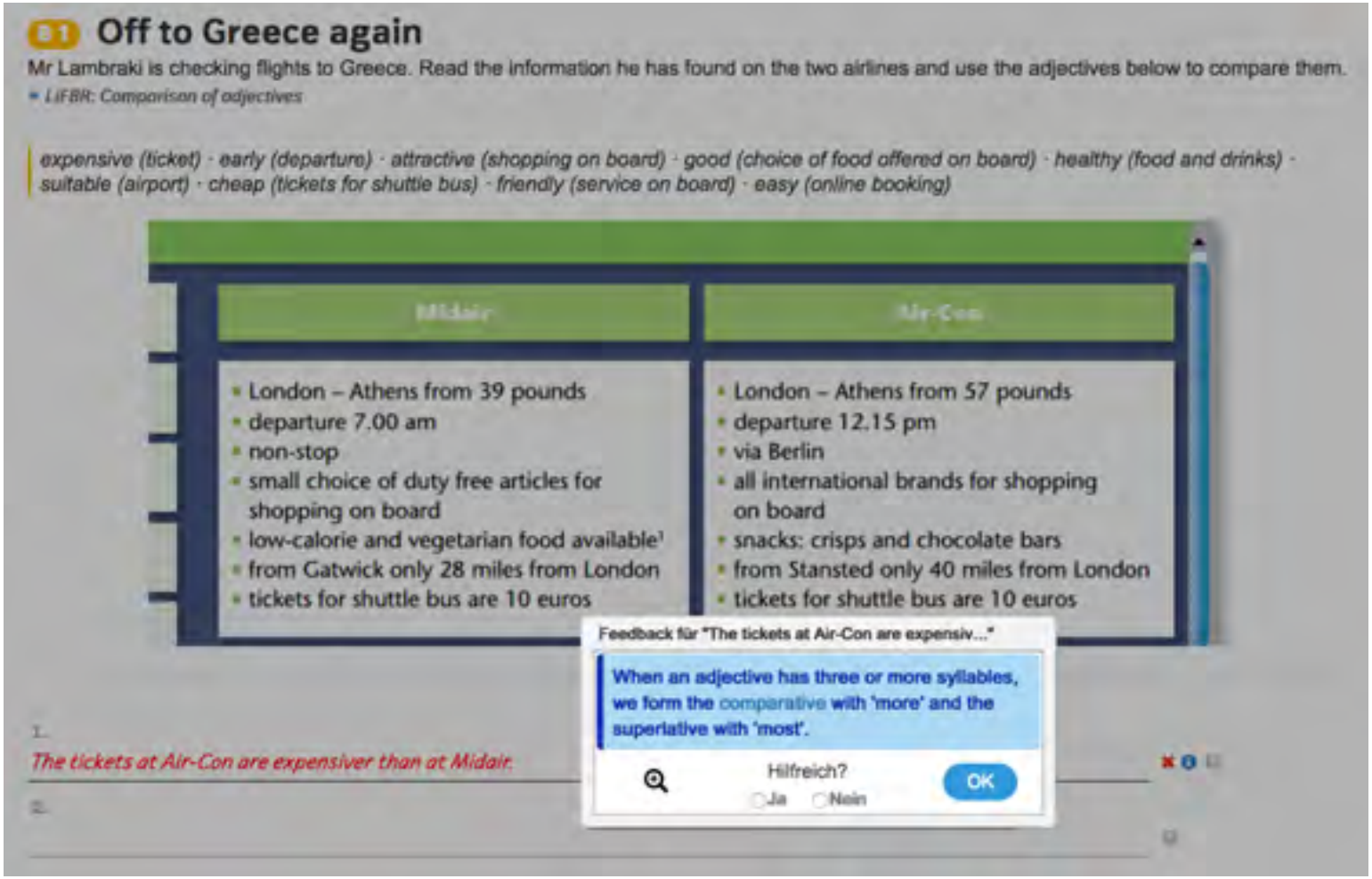
Mr Lambraki is checking flights to Greece. Read the information he has found on the two airlines and use the adjectives below to compare them.

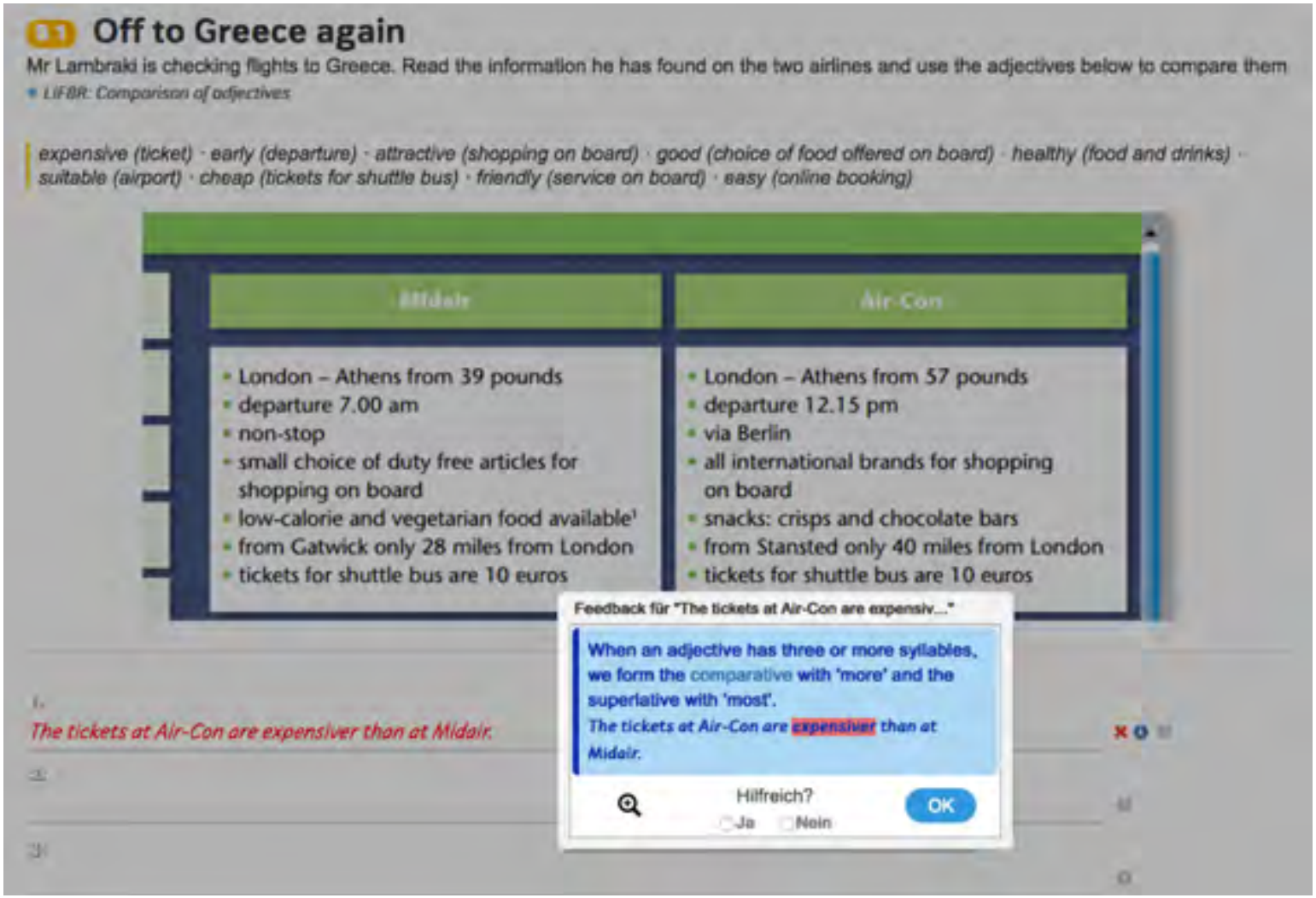
- *LiFBR: Comparison of adjectives*

expensive (ticket) · early (departure) · attractive (shopping on board) · good (choice of food offered on board) · healthy (food and drinks) · suitable (airport) · cheap (tickets for shuttle bus) · friendly (service on board) · easy (online booking)

Midair	Air-Con
• London – Athens from 39 pounds • departure 7.00 am • non-stop • small choice of duty free articles for shopping on board • low-calorie and vegetarian food available[1] • from Gatwick only 28 miles from London • tickets for shuttle bus are 10 euros	• London – Athens from 57 pounds • departure 12.15 pm • via Berlin • all international brands for shopping on board • snacks: crisps and chocolate bars • from Stansted only 40 miles from London • tickets for shuttle bus are 10 euros

1. *The tickets at Air-Con are expensiver than at Midair.*





CYP 3 Reading check: How kayaking changed my life

James is a student at St David's College. Read his report and answer the questions below.

A long, long time ago, I arrived at St David's College ... this is my story...
I wasn't very confident when I first arrived. But I soon found myself in a kayak on the Llangollen canal (...). That was the day kayaking became my life: something I enjoyed, an activity I knew would build my confidence.
I started going to kayaking sessions at weekends and every Wednesday; I loved every minute! I also tried other activities like climbing, mountain biking and sailing. I also tried hill walking but that was rubbish!!!! I started to get really good at kayaking and the outdoor ed teachers[1] helped me develop[2] lots of new skills[3].
I have been lucky enough to go on many expeditions; my first one was in year 10, when I went to Sweden. We sea-kayaked around Tjorn Island (...), and it was an amazing experience which I will never forget.
The same year I went on the Alaska trip (...). The expedition was cold but I still had a really good time. I caught my first fish. Afterwards we ate it, yum[4]!!! We kayaked past glaciers[5] and saw bears[6] and other animals. My kayaking skills and strength[7] were improving, as well as my confidence.
In year 11 I went to Scotland to try white water[8] kayaking, (...) and went down some fantastic rivers. During the expedition I did my first river Eskimo roll[9]. (...) This was a very big moment for me as it greatly helped my confidence in the kayak and after that I got really good.
At the end of year 11 Ian Lloyd Jones suggested that I could do this as a career[10]. From that moment I knew it was all I wanted to do.
In the lower 6th I signed up[11] for (...) an outdoor apprenticeship[12]. (...) I got the opportunity to work with all the year groups on their outdoor ed days and join[13] them on some expeditions; this has been life changing and great fun.
(...)
I also did a very wet year 6 Snowdonia expedition, which was fun until my tent got flooded![14] There was so much water in my tent, I'm sure I could have kayaked in it!!! (...) I'm now a qualified kayaking instructor myself, which is great!!! (...) I am leaving St David's College early to go and work with *Acorn Adventure* in France, teaching kids to canoe[15]. This is what I have wanted for a long time and is the start of my own outdoor ed career.
Thank you St David's for changing my life.
James Oram

[1] outdoor ed teachers – *Sportlehrer für Aktivitäten im Freien (z.B. Kajakfahren etc.)*
[2] develop – *entwickeln*
[3] skills – *Fertigkeiten*
[4] yum – *lecker*
[5] glacier – *Gletscher*
[6] bear – *Bär*
[7] strength – *Stärke*
[8] white water – *Wildwasser*
[9] Eskimo roll – *Eskimorolle (man dreht sich mit dem Kajak so, dass man zeitweise ganz unter Wasser ist)*
[10] career – *Beruf*
[11] sign up – *zu etw anmelden*
[12] apprenticeship – *Ausbildung*
[13] join – *begleiten*
[14] flooded – *überflutet*
[15] canoe – *Kanufahren*

1. How did James feel when he first came to St David's?

James was a student.

CYP 3 **Reading check: How kayaking changed my life**

James is a student at St David's College. Read his report and answer the questions below.

A long, long time ago, I arrived at St David's College ... this is my story ...
I wasn't very confident when I first arrived, but I soon found myself in a kayak on the Llangollen canal (...). That was the day kayaking became my life: something I enjoyed, an activity I knew would build my confidence.
I started going to kayaking sessions at weekends and every Wednesday; I loved every minute! I also tried other activities like climbing, mountain biking and sailing. I also tried hill walking but that was rubbish!!!! I started to get really good at kayaking and the outdoor ed teachers[1] helped me develop[2] lots of new skills[3].
I have been lucky enough to go on many expeditions; my first one was in year 10, when I went to Sweden. We sea-kayaked around Tjorn Island (...), and it was an amazing experience which I will never forget.
The same year I went on the Alaska trip (...). The expedition was cold but I still had a really good time. I caught my first fish. Afterwards we ate it, yum!!!! We kayaked past glaciers[4] and other animals. My kayaki strength[5] were improving, as confidence.
In year 11 I went to Scotland to t kayaking, (...) and went down rivers. During the expedition I di Eskimo roll[6]. (...) This was a very h me as it greatly helped my con

kayak and after that I got really good.
At the end of year 11 Ian Lloyd Jones suggested that I could do this as a career[7]. From that moment I knew it was all I wanted to do.
In the lower 6th I signed up[11] for (...) an outdoor apprenticeship[12]. (...) I got the opportunity to work with all the year groups on their outdoor ed days and join[13] them on some expeditions; this has been life changing and great fun.
(...)
I also did a very wet year 6 Snowdonia expedition, which was fun until my tent got flooded![14] There was so much water in my tent, I'm sure I could have kayaked in it!!! (...)
I'm now a qualified kayaking instructor myself, which is great!!! (...) I am leaving St David's College early to go and work with *Acorn Adventure* in France, teaching kids to canoe[15]. This is what I have wanted for a long time and is the start of my own outdoor ed career.
Thank you St David's for changing my life.
James Oram

Feedback für "James was a student."

There seems to be important information missing in your answer. Please have a look at the highlighted passage in the text.

Hilfreich? Ja Nein OK

1. How did James feel when he first came to St David's?

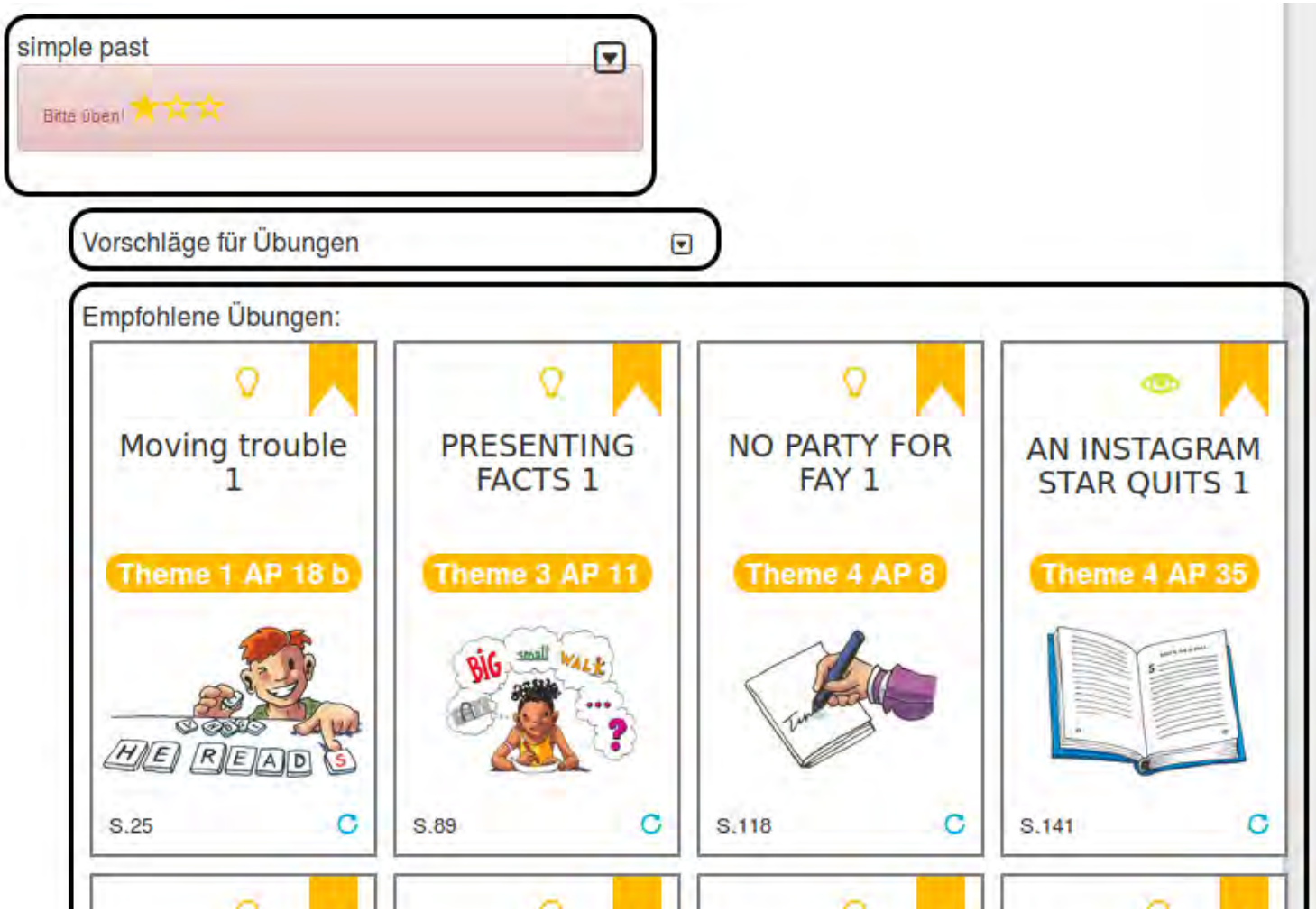
James was a student.

FeedBook characteristics

- FeedBook offers immediate, scaffolded feedback guiding the learner to an understanding of the concepts.
- Analysis of thousands of well-formed and ill-formed variants automatically generated (not manually encoded).
  - Coverage of all grammar topics in 7th grade curriculum
  - 188 different types of feedback on tenses, comparatives, relative & reflexive pronouns, gerunds, passive, conditionals, (in)direct speech
- The FeedBook system also provides
  - learner models, informing student & suggesting next step
  - task models, supporting identification of (in)effective tasks
  - detailed logs of all interactions, i.e., the learning process

Learner model: How am I doing?

Learner model: What should I do next?



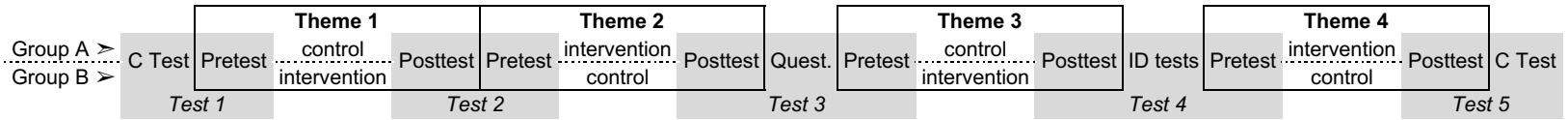
A randomized field trial testing effectiveness

(Meurers et al. 2019)

- Research question:
  - Does immediate formative feedback on form during homework improve learning of English grammar?
- Subjects:
  - Ten 7th grade English classes in Gymnasiums.
- Regular full-year teaching, but FeedBook as workbook.
  - within-class randomization determines who sees specific feedback on which grammar topics

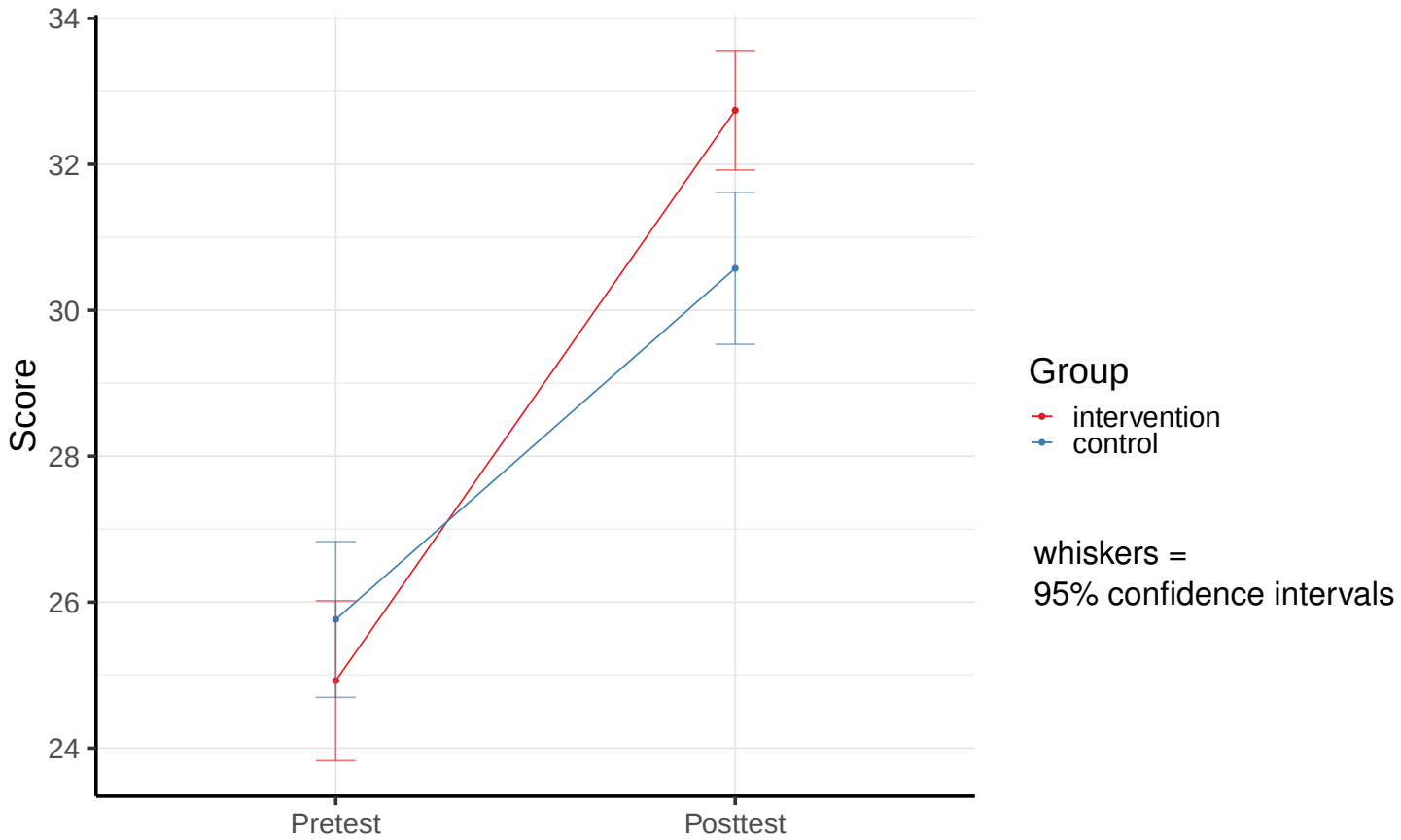
A randomized field trial testing effectiveness

Design of study



- Pre-/Posttests target grammar topics of each theme.
- Analysis here focuses on Theme 2:
  - 205 students completed pre- and posttest
  - grammar topics: comparatives, conditionals, relative clauses
- Results:
  - basic change score analysis
  - mixed effects logistic regression
  - Learning analytics:
    - overall learning process
    - individuals
    - tasks

Results for Theme 2



- Welsh two sample t-test shows significant difference between mean change scores ($p < 0.0001$):
  - 4.81 for control group, 7.82 for intervention group

  → intervention group learned 62% more than control group
- Effect size: Cohen's d = 0.56

Results for Theme 2

Mixed effects logistic regression

- Predict log odds of the binary outcome of each test gap.
  - Fixed effects: test (pre/post), group (interv./control) & their interaction
  - Random intercepts: test items, learners, teachers/classes

	Estimate	Std. Error
(Intercept) = *pretest control*	0.73***	(0.15)
TestPosttest	0.68***	(0.05)
GroupIntervention	−0.07	(0.09)
TestPosttest:GroupIntervention	0.43***	(0.08)
Var: Learner (Intercept)	0.30	
Var: Item (Intercept)	0.52	
Var: Teacher (Intercept)	0.05	

***$p < 0.001$

Tracking learning “in the wild”

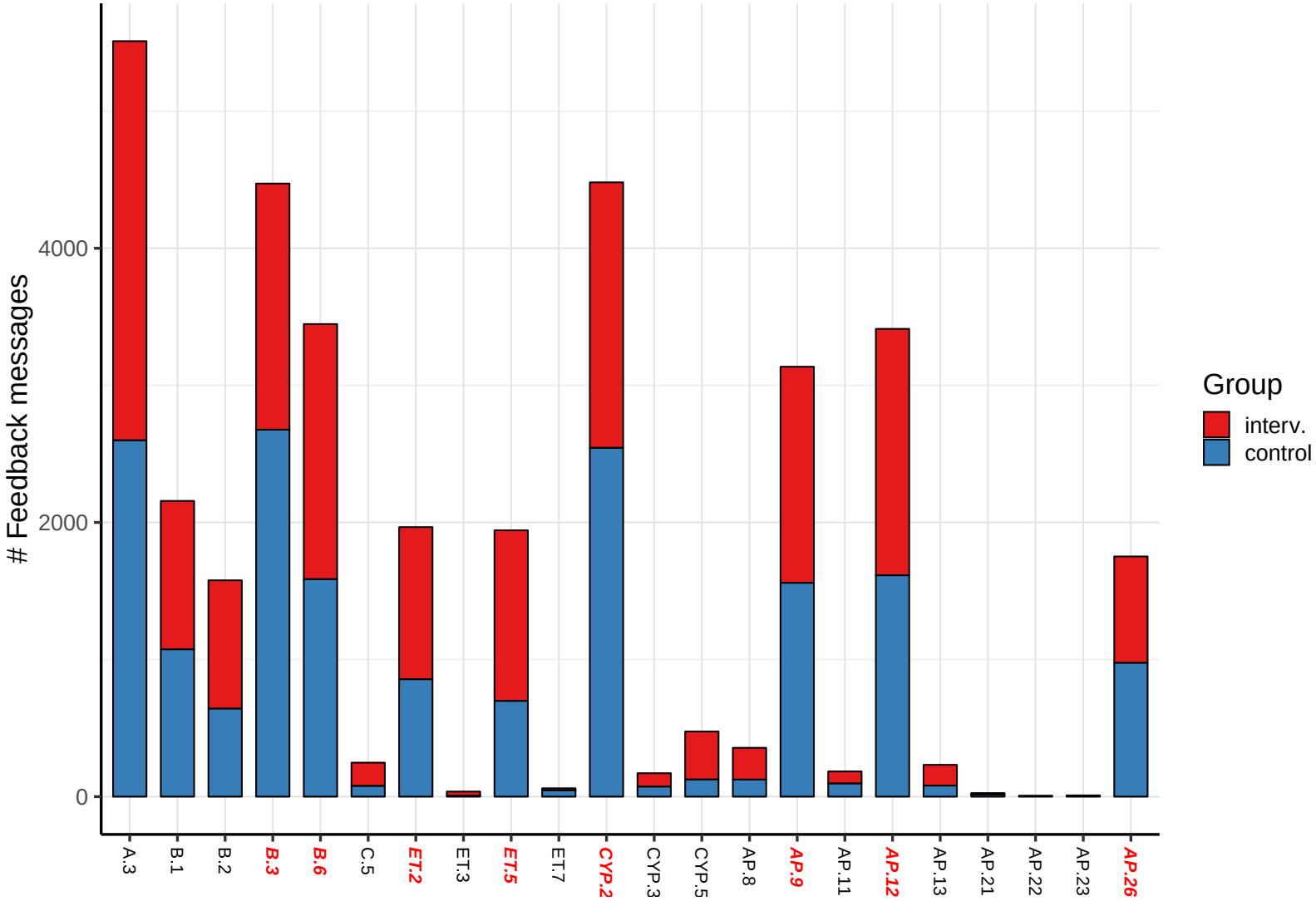
- Full yearlong integration into regular classes
  - five different schools with different profiles and
  - different teachers following their regular teaching

  → high ecological validity
- But also loss of control and focus:
  - Which homework did the teachers assign?
    - We only asked for inclusion of 2–3 exercises/topic.
  - Which homework did the children actually do?
  - How often did students receive what kind of feedback?
  - Who showed uptake for the tested language constructs?

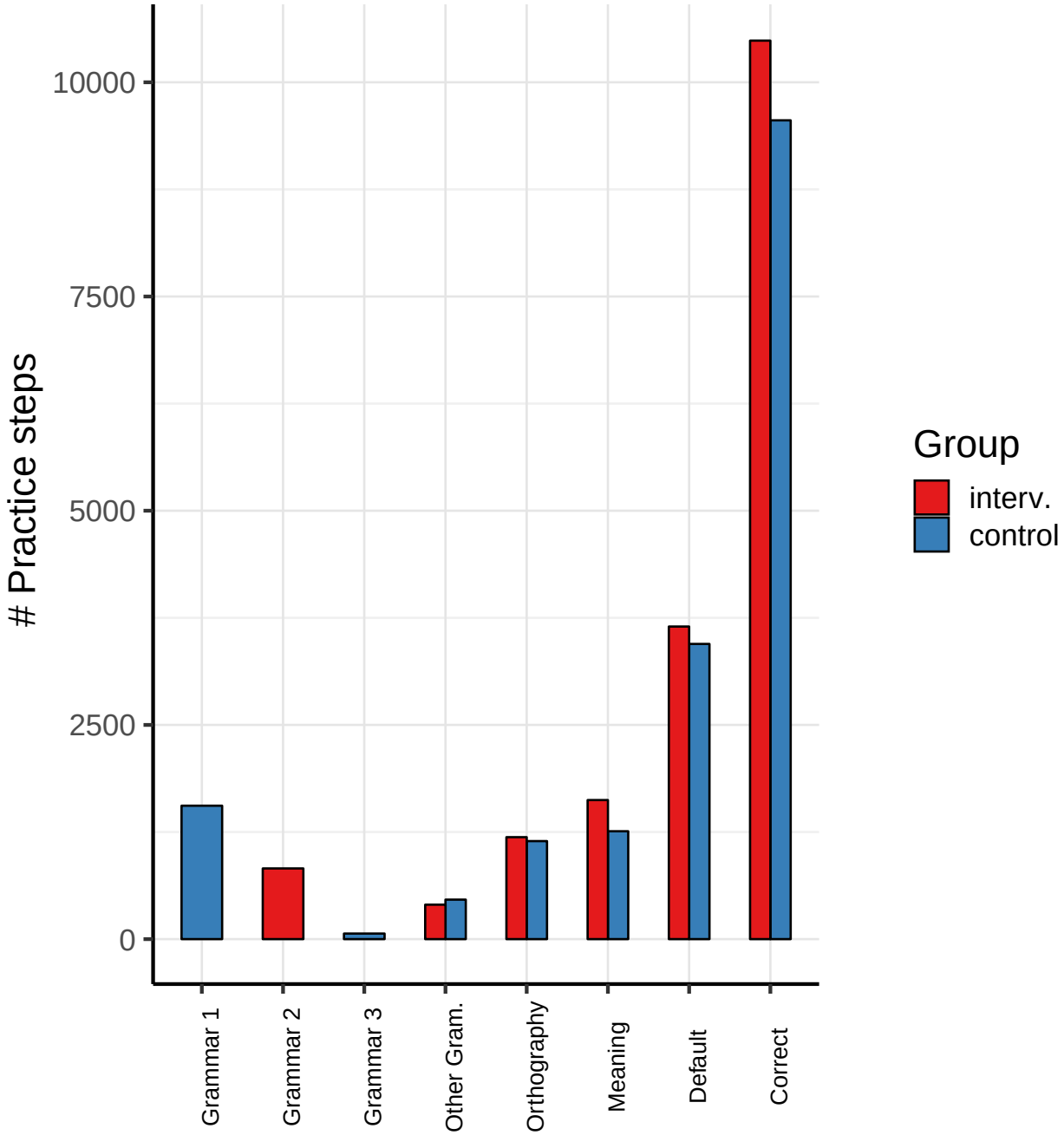
→ We can’t control it, but we can log all interactions to support post-hoc analysis of what happened!

- Combined with additional web-based individual tests:
  - working memory capacity (OSPAN), MLAT-V, C-Test, typing ability, questionnaires, . . .

Exercises worked on & amount of feedback seen

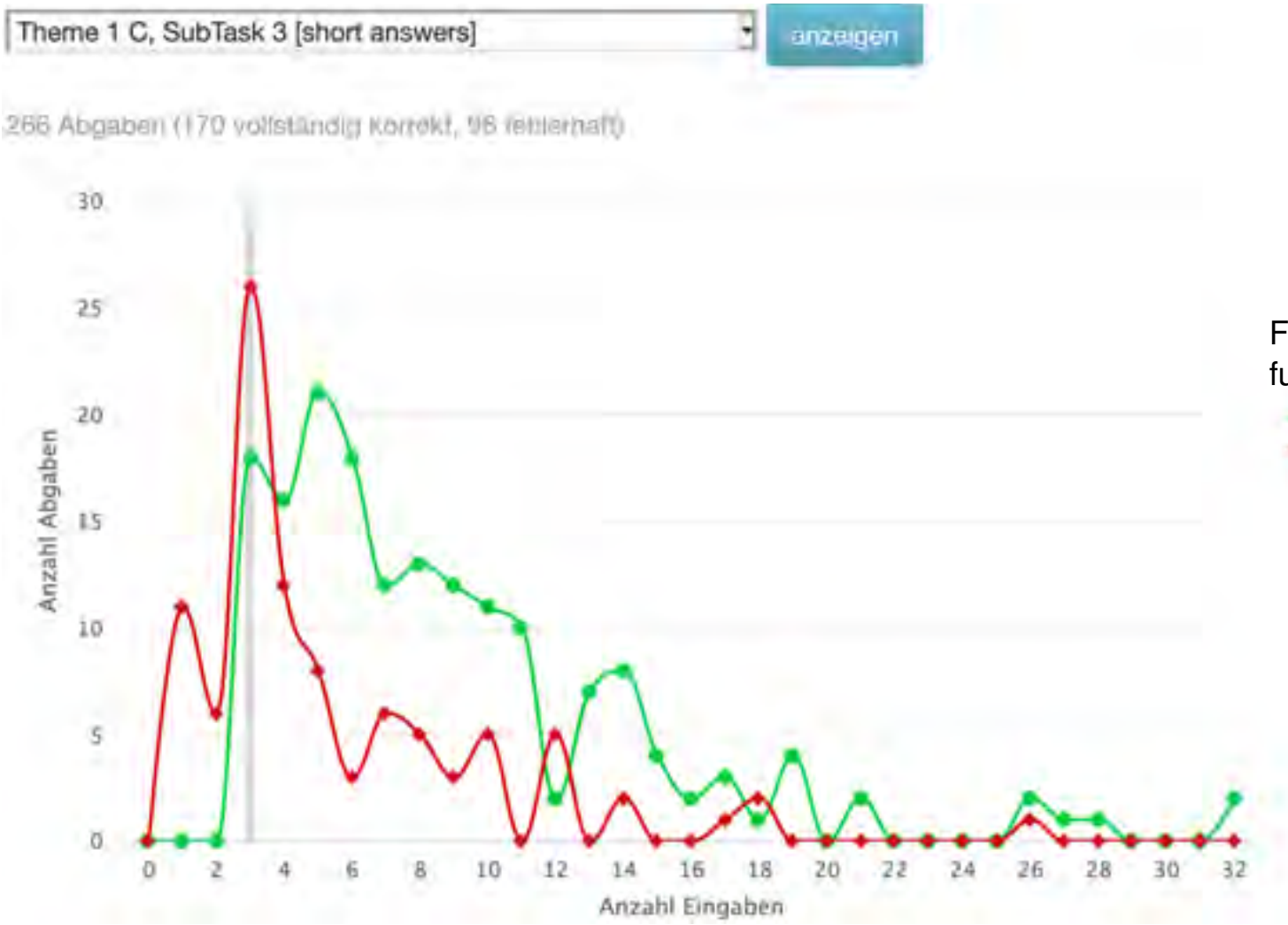


Types of feedback seen



Log data: Analyzing task interactions

Which tasks support learning for this population?

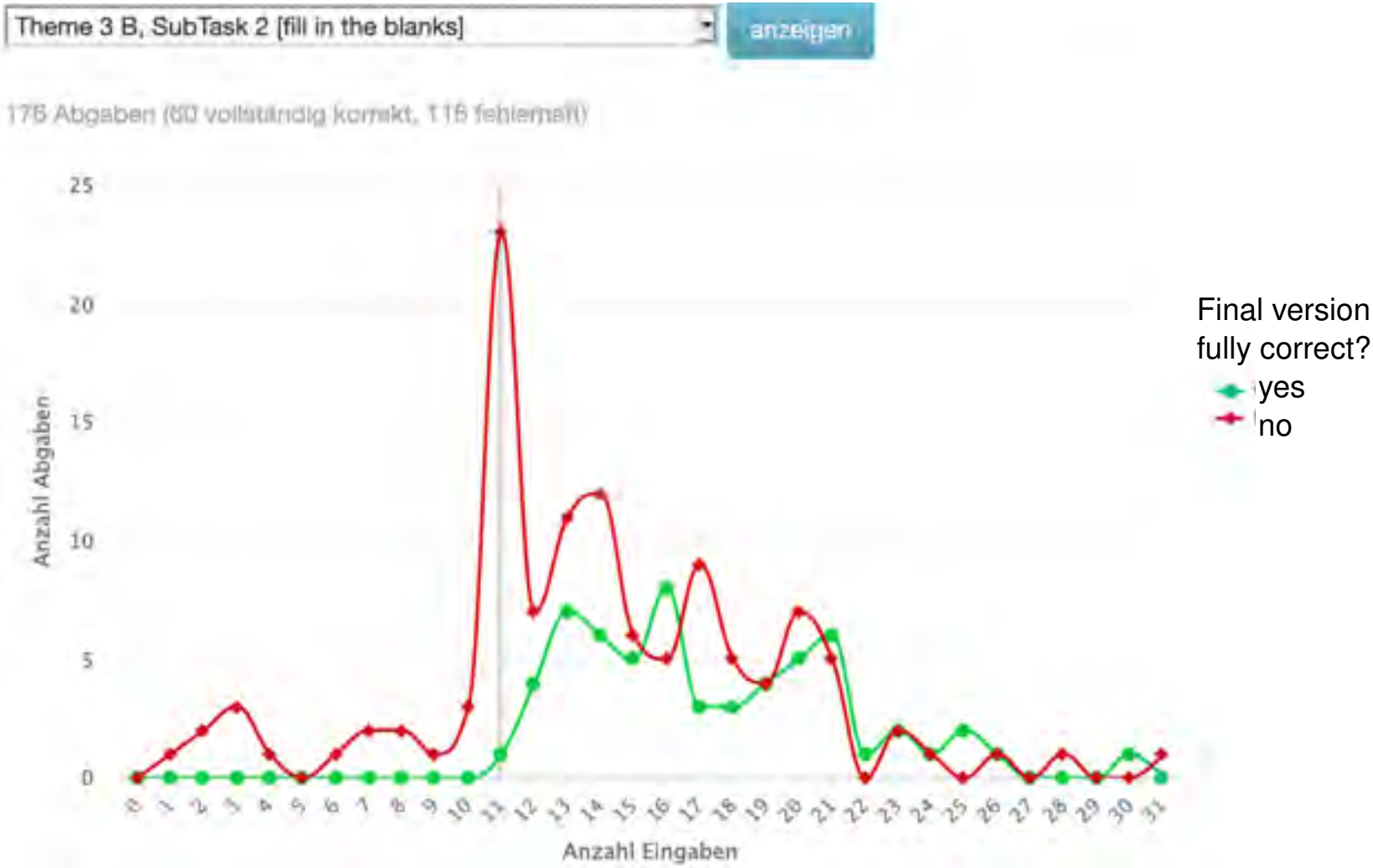


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

LEAD
Graduate School

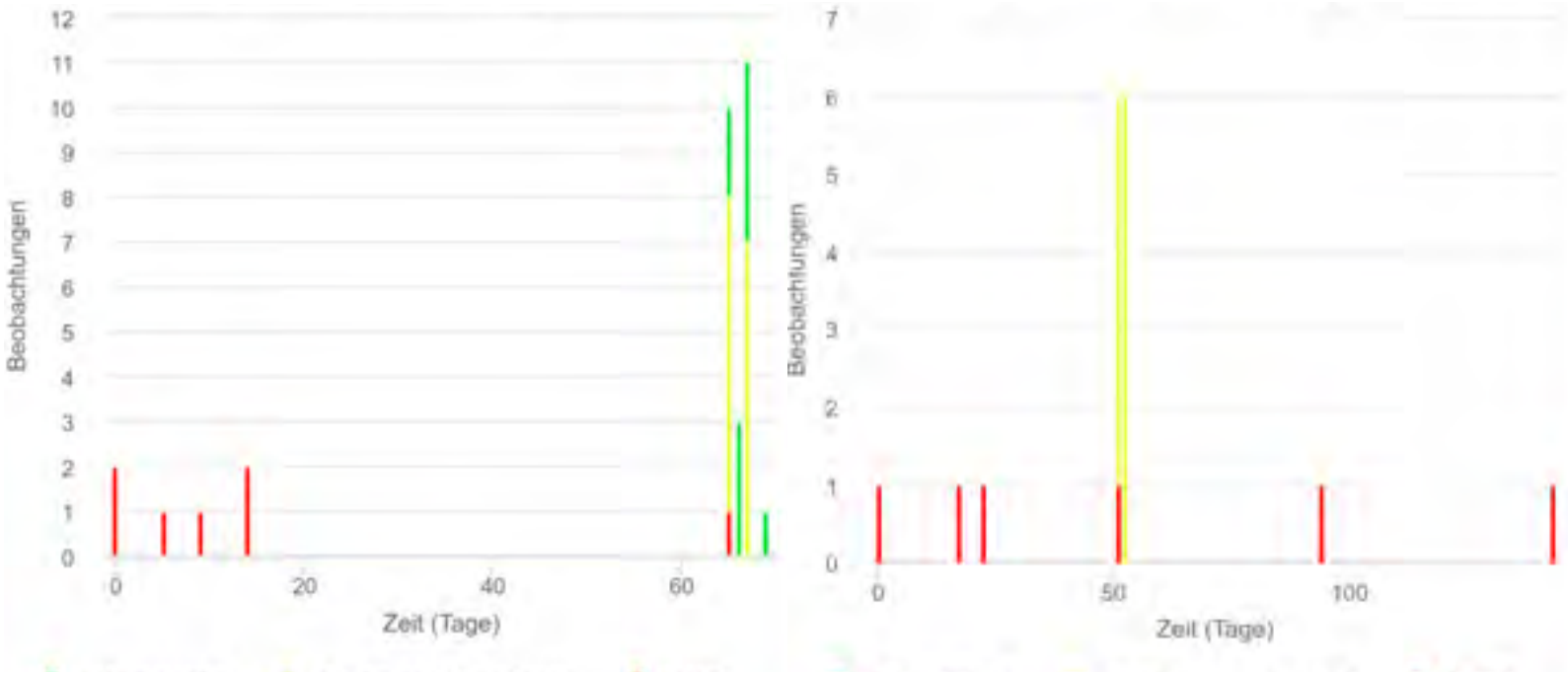
Log data: Analyzing task interactions

. . . and which less so?



Log data: Individual uptake sequences

Analysis of two learners on comparatives:



- Under what conditions can we determine proficiency based on longitudinal logs of the process of completing exercises?
  - When is this more reliable? more valid?
  - Determining language proficiency arguably is not about unique top performances, as in sports.

Conclusions (I)

- Computational linguistic methods support the automatic analysis of learner language:
  - specific aspects of the language system
  - general linguistic complexity

    Relevant for large-scale data analysis and supporting immediate learner interaction.
- CL analysis traditionally starts from the language data, e.g., providing analyses of linguistic complexity.
  - But task factors need to be considered for valid analyses.

  → Develop parametrizable models of task complexity
    - simplify item construction and ensure validity of analyses
    - support dynamic difficulty adjustment: adapt task complexity to individual learner's current ability
    - machine learned models more likely than full theories

Conclusions (II)

- For tasks providing top-down guidance, CL analysis can start from the task specification, such as target answers.
  - Generate well-formed and ill-formed variability supported by the task, then flexibly match to learner answer.
  - For automatic assessment of reading comprehension answers, the question can help focus the analysis (Ziai & Meurers 2014, 2018).
- We illustrated task-based processing with the FeedBook
  - first randomized field study of Intelligent Language Tutoring System (ILTS) fully embedded in German schools
  - results show significant improvement for children receiving specific grammar feedback
- ILTS provide rich data on learning process, for which methods to interpret such data need to be developed.
- Lots of opportunity for collaboration between Computational Linguistics and Language Testing!

References

Alexopoulou, T., J. Geertzen, A. Korhonen & D. Meurers (2015). Exploring big educational learner corpora for SLA research: perspectives on relative clauses. *International Journal of Learner Corpus Research* 1(1), 96–129.

Alexopoulou, T., M. Michel, A. Murakami & D. Meurers (2017). Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. *Language Learning* 67, 181–209. URL https://doi.org/10.1111/lang.12232.

Beinborn, L. (2016). Predicting and manipulating the difficulty of text completion exercises for language learning. Ph.D. thesis, Department of Computer Science, Technische Universität Darmstadt. URL http://tuprints.ulb.tu-darmstadt.de/5647/.

Berkling, K., J. Fay, M. Ghayoomi, K. Hein, R. Lavalley, L. Linhuber & S. Stüker (2014). A Database of Freely Written Texts of German School Students for the Purpose of Automatic Spelling Error Classification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 1212–1217. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/255_Paper.pdf. Available via ACL from: https://catalog.ldc.upenn.edu/LDC2015T22.

Chen, X. (2018). Automatic Analysis of Linguistic Complexity and Its Application in Language Learning Research. Ph.D. thesis, Eberhard Karls Universität Tübingen Germany. URL http://hdl.handle.net/10900/85888.

Chen, X. & D. Meurers (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*. Osaka, Japan: COLING, pp. 113–119.

Chen, X. & D. Meurers (2017a). Challenging Learners in Their Individual Zone of Proximal Development Using Pedagogic Developmental Benchmarks of Syntactic Complexity. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017*. Gothenburg, Sweden: ACL, Linköping Electronic Conference Proceedings 134, pp. 8–17. URL http://aclweb.org/anthology/W17-0302.pdf.

Chen, X. & D. Meurers (2017b). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading* 41(3), 486–510.

Chen, X. & D. Meurers (2019). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer-Assisted Language Learning* URL https://doi.org/10.1080/09588221.2018.1527358.

Daroczy, G., M. Wolska, W. D. Meurers & H.-C. Nuerk (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology* 6(348). URL http://www.frontiersin.org/developmental_psychology/10.3389/fpsyg.2015.00348/abstract.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford, UK: Oxford University Press.

Geertzen, J., T. Alexopoulou & A. Korhonen (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum (SLRF)*. Cascadilla Press. URL http://purl.org/icall/efcamdat.

Karges, K., T. Studer & E. Wiedenkeller (in press). On the way to a new multilingual learner corpus of foreign language learning in school: observations about task variations. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (eds.), *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Louvain-La-Neuve: Presses Universitaires de Louvain.

Kyle, K. (2016). Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Ph.D. thesis, Georgia State University. URL http://scholarworks.gsu.edu/alesl_diss/35.

Lu, X. (2014). *Computational Methods for Corpus Annotation and Analysis*. Springer.

McNamara, D. S., A. C. Graesser, P. M. McCarthy & Z. Cai (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Meurers, D. (2012). Natural Language Processing and Language Learning. In C. A. Chapelle (ed.), *Encyclopedia of Applied Linguistics*, Oxford: Wiley, pp. 4193–4205. URL http://purl.org/dm/papers/meurers-12.html.

Meurers, D. (2015). Learner Corpora and Natural Language Processing. In S. Granger, G. Gilquin & F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, pp. 537–566. http://purl.org/dm/papers/meurers-15.html.

Meurers, D., K. De Kuthy, F. Nuxoll, B. Rudzewitz & R. Ziai (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics* 39. URL http://purl.org/dm/papers/Meurers.DeKuthy.ea-19.pdf. To appear.

Meurers, D. & M. Dickinson (2017). Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration with Computational Linguistics. *Language Learning* 67(2).

Meurers, D., N. Ott & R. Ziai (2010). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In *Pre-Proceedings of Linguistic Evidence*. Tübingen, pp. 214–217. URL http://purl.org/dm/papers/meurers-ott-ziai-10.html.

Michel, M., A. Murakami, T. Alexopoulou & D. Meurers (accepted). Effects of task type on morphosyntactic complexity across proficiency: Evidence from a large learner corpus of A1 to C2 writings. *Instructed Second Language Acquisition* .

Miura, S. (1998). Hiroshima English Learners' Corpus: English learner No. 2 (English I & English II). Department of English Language Education, Hiroshima University. http://purl.org/icall/helc.

Ott, N., R. Ziai & D. Meurers (2012). Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context. In T. Schmidt & K. Wörner (eds.), *Multilingual Corpora and Multilingual Corpus Analysis*, Amsterdam: Benjamins, Hamburg Studies in Multilingualism (HSM), pp. 47–69. URL https://benjamins.com/#catalog/books/hsm.14.05ott.

Pandarova, I., T. Schmidt, J. Hartig, A. Boubekki, R. D. Jones & U. Brefeld (2019). Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education* To appear.

Quixal, M. & D. Meurers (2016). How can writing tasks be characterized in a way serving pedagogical goals and automatic analysis needs? *CALICO Journal* 33, 19–48. URL http://purl.org/dm/papers/Quixal.Meurers-16.html.

Rescher, N. (1998). *Complexity: A philosophical overview*. London: Transaction Publishers.

Robinson, P. (1995). Attention, memory, and the “noticing” hypothesis. *Language Learning* 45(2), 283–331.

Rudzewitz, B., R. Ziai, K. De Kuthy & D. Meurers (2017). Developing a web-based workbook for English supporting the interaction of students and teachers. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition*. URL http://aclweb.org/anthology/W17-0305.pdf.

Rudzewitz, B., R. Ziai, K. De Kuthy, V. Möller, F. Nuxoll & D. Meurers (2018). Generating Feedback for English Foreign Language Exercises. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. ACL, pp. 127–136. URL http://aclweb.org/anthology/W18-0513.pdf.

Skehan, P. (1989). *Individual Differences in Second Language Learning*. Edward Arnold.

Skehan, P. (1998). *A Cognitive Approach to Learning Language*. Oxford University Press.

Svetashova, Y. (2015). C-Test Item Difficulty Prediction: Exploring the Linguistic Characteristics of C-Tests Using Machine Learning. Master’s thesis in computational linguistics, Department of Linguistics, University of Tübingen. 186pp.

Vajjala, S. (2015). Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications. Ph.D. thesis, University of Tübingen.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. Montréal, Canada: ACL, pp. 163–173. URL http://aclweb.org/anthology/W12-2019.pdf.

Vajjala, S. & D. Meurers (2013). On The Applicability of Readability Models to Web Texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. pp. 59–68.

Vajjala, S. & D. Meurers (2014a). Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. ACL, Gothenburg, Sweden: ACL, pp. 288–297. URL https://aclweb.org/anthology/E14-1031.pdf.

Vajjala, S. & D. Meurers (2014b). Exploring Measures of "Readability" for Spoken Language: Analyzing linguistic features of subtitles to identify age-specific TV programs. In *Proceedings of the Third Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Gothenburg, Sweden: ACL, pp. 21–29.

Vajjala, S. & D. Meurers (2014c). Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification* 165(2), 142–222.

Vyatkina, N. (2012). The Development of Second Language Writing Complexity in Groups and Individuals: A Longitudinal Learner Corpus Study. *The Modern Language Journal* 96(4), 576–598. URL https://doi.org/10.1111/j.1540-4781.2012.01401.x.

Wang, C. & M. Wang (2015). Effect of alignment on L2 written production. *Applied Linguistics* 36(5), 503–526.

Weiss, Z. & D. Meurers (2018). Modeling the Readability of German Targeting Adults and Children: An Empirically Broad Analysis and its Cross-Corpus Validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*. Santa Fe, New Mexico, USA: International Committee on Computational Linguistic. URL https://www.aclweb.org/anthology/C18-1026.

Weiss, Z. & D. Meurers (in press). Broad Linguistic Modeling is Beneficial for German L2 Proficiency Assessment. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (eds.), *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Louvain-La-Neuve: Presses Universitaires de Louvain.

Wisniewski, K. (2017). Empirical Learner Language and the Levels of the Common European Framework of Reference. *Language Learning* 67(S1), 232–253. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12223.

Ziai, R. & D. Meurers (2014). Focus Annotation in Reading Comprehension Data. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII, 2014)*. COLING, Dublin, Ireland: ACL, pp. 159–168. URL http://aclweb.org/anthology/W14-4922.pdf.

Ziai, R. & D. Meurers (2018). Automatic Focus Annotation: Bringing Formal Pragmatics Alive in Analyzing the Information Structure of Authentic Data. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. New Orleans, LA: ACL, pp. 117–128. URL http://aclweb.org/anthology/N18-1011.pdf.

Ziai, R., B. Rudzewitz, K. De Kuthy, F. Nuxoll & D. Meurers (2018). Feedback Strategies for Form and Meaning in a Real-life Language Tutoring System. In *Proceedings of the 7th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)*. ACL, pp. 91–98. URL http://aclweb.org/anthology/W18-7110.

Short answer assessment

Task information supporting analysis

- Automatic short answer assessment cannot rely on rich lexical distributional information, as essay scoring does.
- For reading comprehension questions, we can make use of the rich task structure:
  - the question asked
  - the information sources in the text, and based on these,
  - the identification of givenness and focus in the answer.

Short answer assessment

Data used

- CREG (Meurers, Ott & Ziai 2010; Ott, Ziai & Meurers 2012)
  - reading comprehension corpus ≈ 35,000 student answers
  - plus corresponding target answers, questions and reading texts
  - every student answer was classified as correct or incorrect by two teachers

Short answer assessment

Illustrating alignment-based analysis

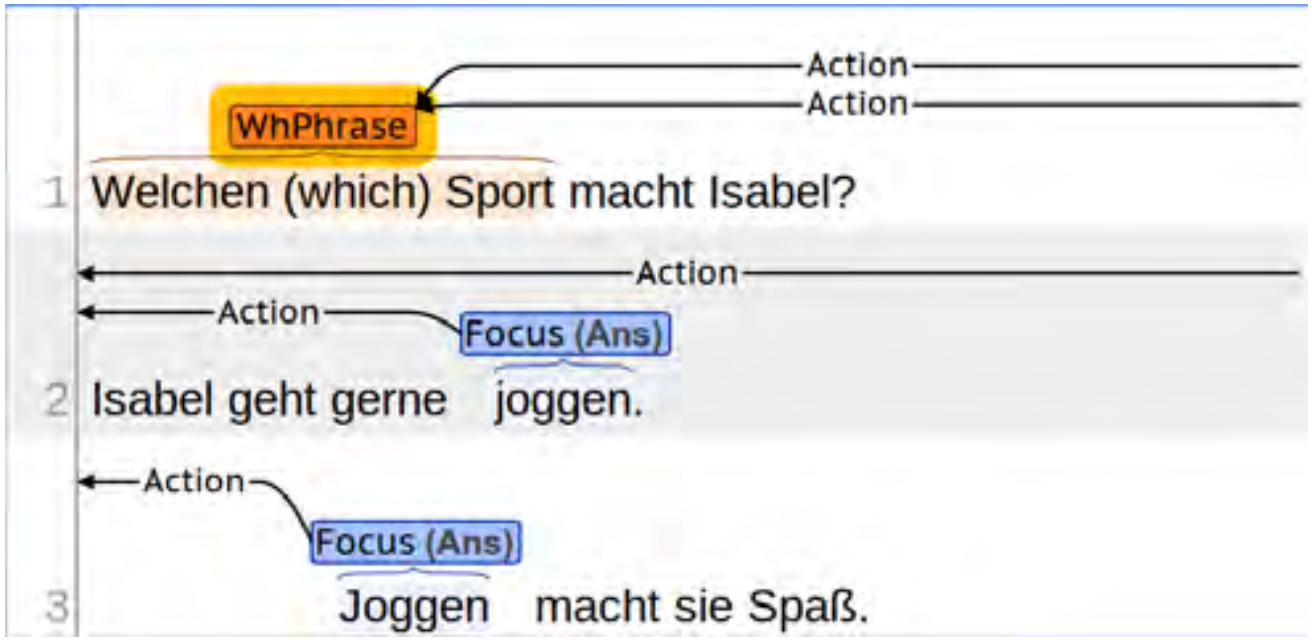
Q: **Was sind die Kritikpunkte, die Leute über Hamburg äußern?**
'What are the objections people have about Hamburg?'

TA: **Der Gestank von Fisch und Schiffsdiesel an den Kais .**
The stink of fish and fuel at the quays .

SemType: Der Gestank
Spelling: von Fisch
Token: und Schiffsdiesel
Chunk: an den Kais

SA: **Der Geruch zon Fish und Schiffsdiesel beim Hafen .**
The smell of_{err} $fish_{err}$ and fuel at the port .

Alignment restricted to focused constituent



Q: ‘**Which sport** does Isabel do?’

TA: ‘She likes to go $[\![\text{jogging}]\!]_F$.’

SA: ‘$[\![\text{Jogging}]\!]_F$ is fun for her.’

Short answer assessment

Analysis of Focus improves results

	Standard	With Focus	
CREG-1032	85.9%	88.6%	+2.7%
CREG-2155	82.1%	85.1%	+3.0%
Overall	83.2%	85.6%	+2.4%

Complexity alignment

Relating linguistic complexity in writing & input

- Can we measure the impact of linguistic complexity of the input on the learner's output directly?
- Data: Wang & Wang (2015) asked English learners to
  (i) continue writing a Chinese text: **baseline writing**
  (ii) continue writing an English text: **continuation writing**

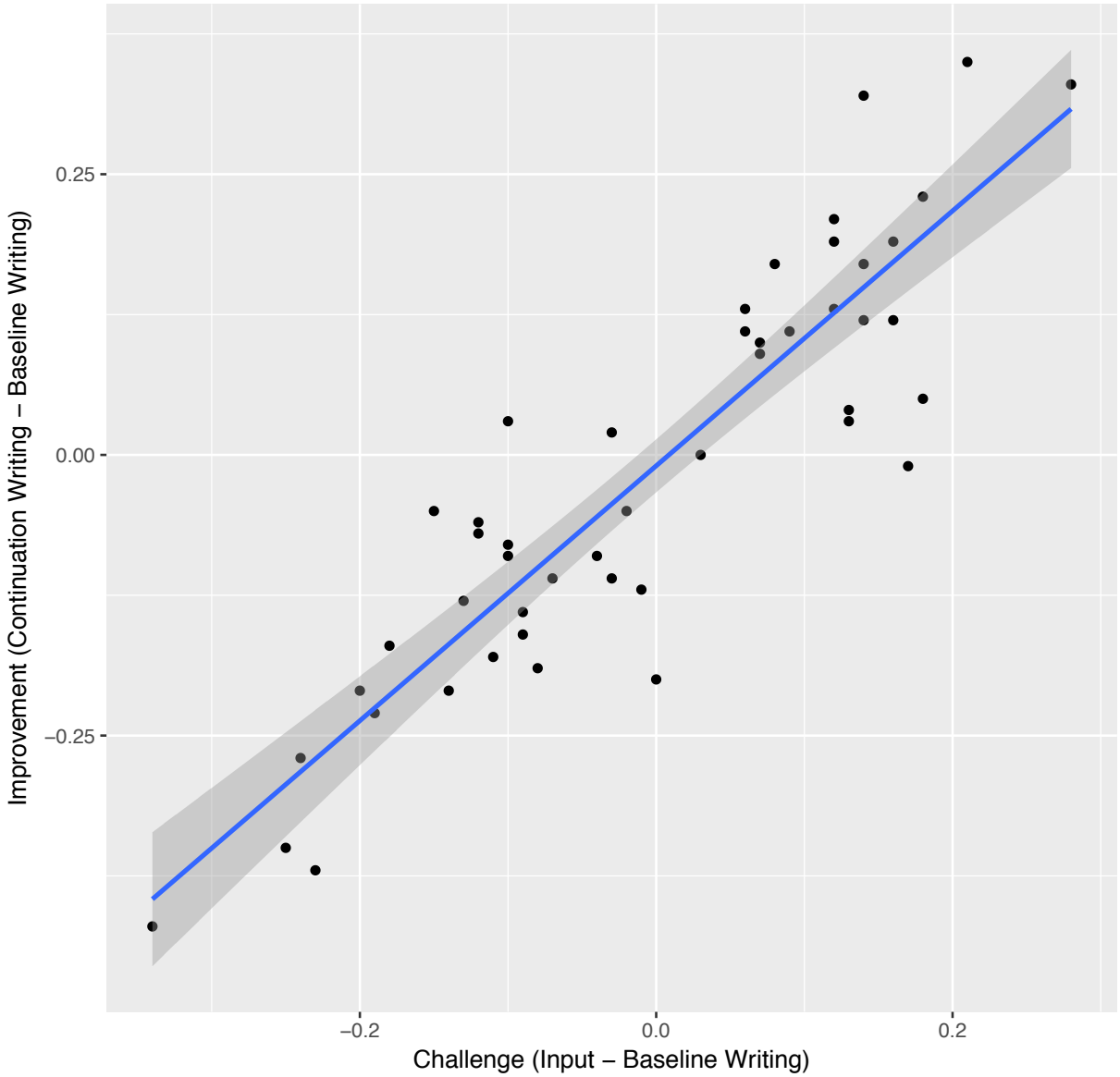
  and found fewer errors in (ii) = alignment with the input
- Alignment with input also measurable for ling. complexity?
- Chen & Meurers (2019) reanalyzes the data & compares:

  Challenge
  Complexity(English input text) – Complexity(baseline writing)

  Improvement
  Complexity(continuation writing) – Complexity(baseline writing)

Alignment of output to the input

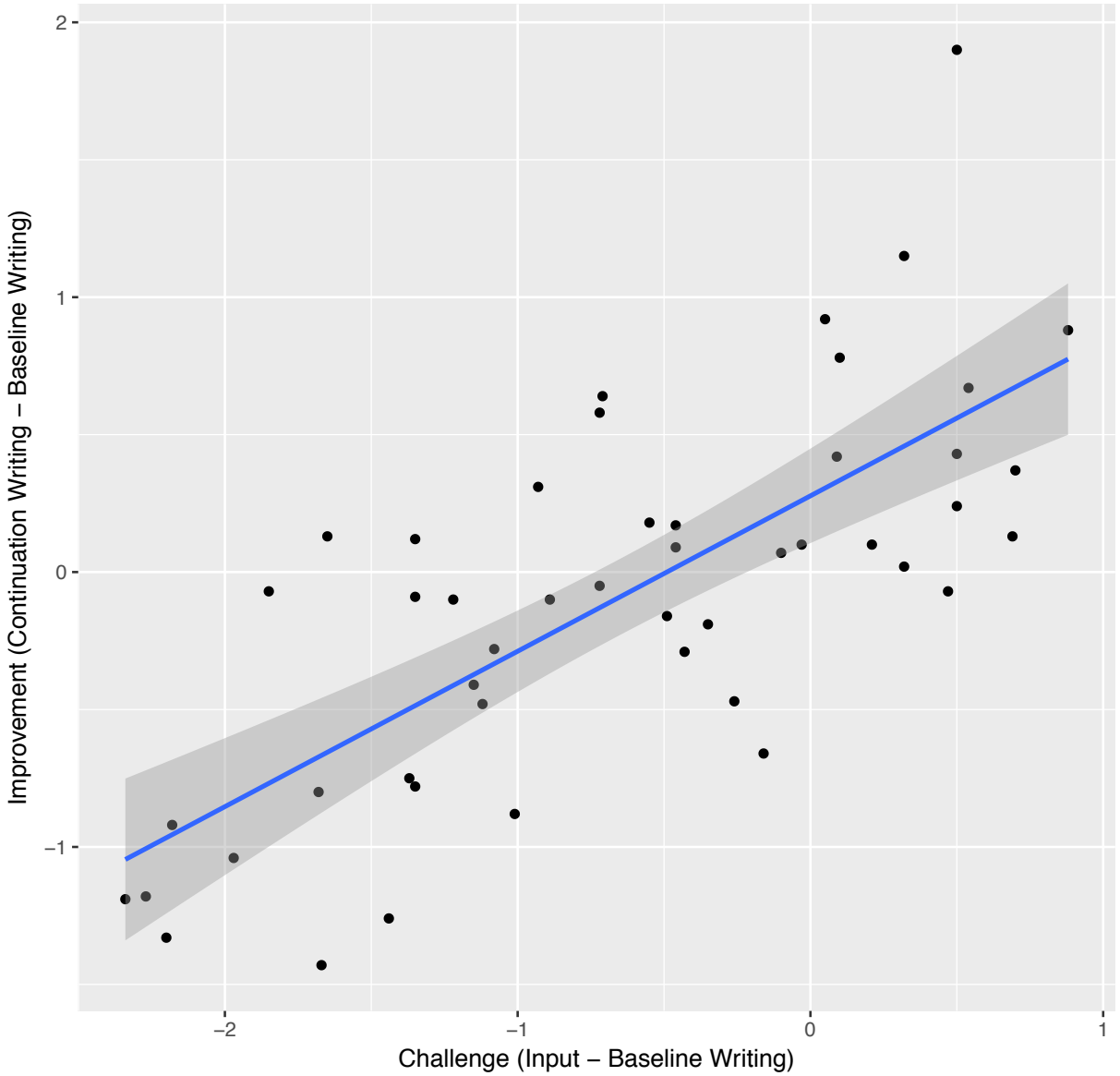
Average length of NP



$(r^2 = 81\%)$

Alignment of output to input

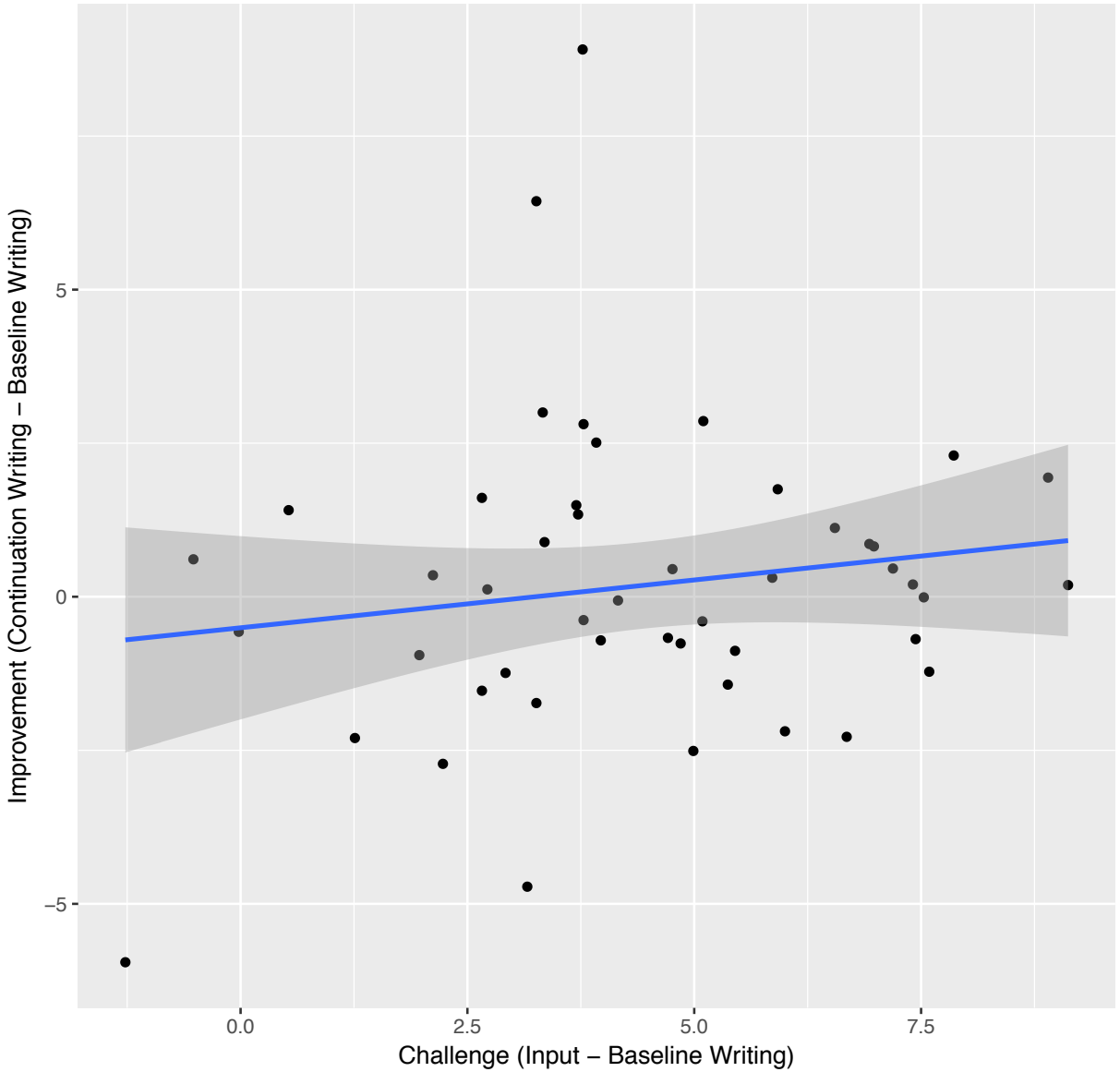
Lexical diversity (MTLD)



$(r^2 = 54\%)$

Alignment of output to input

Cohesion: Global noun overlap



$(r^2 = 2\%)$

Exploring alignment

- For 489/559 complexity measures, the result is significant.
- Challenge explains most variance in improvement for:
  - avg. length of noun phrases (81%)
  - avg. number of verbs in past tense (78%)
- Measures showing little effect:
  - modifier variation (28%)
  - global noun overlap (2%)
- We find alignment to input both below and above level of *i*
  - More Challenge needed to see where leveling off above +1
  - Longitudinal observation needed to see if alignment results in learning, incrementally advancing abilities

$\Rightarrow$ support selection of individually challenging texts

- Syntactic Benchmark (SyB) (Chen & Meurers 2017a, 2019)

SyB: Syntactic Benchmark (http://complexityweb.org)