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

As shown in Figure 7, the use of prepositions consistently increases from early beginner to advanced levels while the accuracy shown in Figure 9 suggests a U-shape pattern, dropping at level 5 to then increase until late intermediate (level 13). In other words, the initial increase in use gives rise to a
Complexity development in EFCamDat
Syntactic elaborateness: subordination

![Graph showing complexity development in EFCamDat](image-url)
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

The graph shows the metric MTLD (Metric for Lexical Diversity) across different levels of CEFR (Common European Framework of Reference for Languages) and Enlishtown Levels, as well as by task type (Desc = Description, Nar = Narrative, Prof = Proficiency). The data points and error bars indicate the variability in MTLD scores across these levels and task types.
Complexity and task effects in EFCamDat
Syntactic elaborateness: subordination

Subordinate Clause per T-unit

<table>
<thead>
<tr>
<th>CEFR</th>
<th>Englishtown Level</th>
<th>By Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.10</td>
<td>Nar 6:1</td>
</tr>
<tr>
<td>A2</td>
<td>0.15</td>
<td>Nar 7:4</td>
</tr>
<tr>
<td>B1</td>
<td>0.20</td>
<td>Desc 6:7</td>
</tr>
<tr>
<td>B2</td>
<td>0.25</td>
<td>Desc 7:7</td>
</tr>
<tr>
<td>C1</td>
<td></td>
<td>Nar 6:1</td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td>Nar 7:4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prof 6:4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prof 7:3</td>
</tr>
</tbody>
</table>
Zooming in on tasks

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

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.

⇒ Valid interpretation requires explicit task / language context.
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.
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.

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

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 tried 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
**Introduction**

**Linguistic complexity: data-driven analysis**

Definition & empirical basis
- Analyzing development in large-scale data
- Task effects on complexity measures of development
- Manipulating task complexity
- Interpreting learner language

**Task-driven analysis**

The FeedBook

FeedBook characteristics
- Testing effectiveness in a randomized field trial
- Results of the field trial

**Conclusions**

**References**

**Appendix**

Short answer assessment
- Complexity alignment
Introduction

Linguistic complexity: data-driven analysis

Definition & empirical basis
Analyzing development in large-scale data
Task effects on complexity measures of development
Manipulating task complexity
Interpreting learner language

Task-driven analysis

The FeedBook

FeedBook characteristics
Testing effectiveness in a randomized field trial
Results of the field trial

Conclusions

References

Appendix

Short answer assessment
Complexity alignment
Introduction

Linguistic complexity: data-driven analysis
- Definition & empirical basis
- Analyzing development in large-scale data
- Task effects on complexity measures of development
- Manipulating task complexity
- Interpreting learner language

Task-driven analysis

The FeedBook
- FeedBook characteristics
- Testing effectiveness in a randomized field trial
- Results of the field trial

Conclusions

References

Appendix
- Short answer assessment
- Complexity alignment

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.

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
   We are talking about something that was happening in the past at the same time as something else. An expression like 'while' shows that this was happening for a longer time, so we use the past progressive.
Introduction

Linguistic complexity: data-driven analysis
Definition & empirical basis
Analyzing development in large-scale data
Task effects on complexity measures of development
Manipulating task complexity
Interpreting learner language

Task-driven analysis
The FeedBook
FeedBook characteristics
Testing effectiveness in a randomized field trial
Results of the field trial

Conclusions

References

Appendix
Short answer assessment
Complexity alignment
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.

- 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)

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

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

- LiFB: 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)**

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

---

The tickets at Air-Con are more expensive than at Midair.

Feedback for "The tickets at Air-Con are more expensive than at Midair."

When an adjective has three or more syllables, we form the comparative with 'more' and the superlative with 'most'.

Hilfreich? [Ja / Nein]
Introduction

Linguistic complexity: data-driven analysis

Definition & empirical basis
Analyzing development in large-scale data
Task effects on complexity measures of development
Manipulating task complexity
Interpreting learner language

Task-driven analysis

The FeedBook
FeedBook characteristics
Testing effectiveness in a randomized field trial
Results of the field trial

Conclusions

References

Appendix

Short answer assessment
Complexity alignment
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 helped me develop lots of new skills!
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 and saw bears and other animals. My kayaking skills and strength were improving, as well as my confidence.
In year 11 I went to Scotland to try white water kayaking (...). I also went on some fascinating rivers. During the expedition I did my first river Eschimo roll! (...) 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. From that moment I knew it was all I wanted to do. In the lower sixth I signed up for (...) an outdoor apprenticeship (...). I got the opportunity to work with all the year groups on their outdoor ed days and join 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! 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'm leaving St David's College early to go and work with Acorn Adventure in France, teaching kids to canoe. 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 Cram

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

James was a student.
### Introduction

#### Linguistic complexity: data-driven analysis

- **Definition & empirical basis**
- **Analyzing development in large-scale data**
- **Task effects on complexity measures of development**
- **Manipulating task complexity**
- **Interpreting learner language**

### Task-driven analysis

#### The FeedBook

- **FeedBook characteristics**
- **Testing effectiveness in a randomized field trial**

### Conclusions

### References

### Appendix

- **Short answer assessment**
- **Complexity alignment**
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

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept) = pretest control</td>
<td>0.73***</td>
</tr>
<tr>
<td>TestPosttest</td>
<td>0.68***</td>
</tr>
<tr>
<td>GroupIntervention</td>
<td>−0.07</td>
</tr>
<tr>
<td>TestPosttest:GroupIntervention</td>
<td>0.43***</td>
</tr>
</tbody>
</table>

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 (OSPA), MLAT-V, C-Test, typing ability, questionnaires, . . .
Exercises worked on & amount of feedback seen

Exercise # Feedback messages
Group
interv.
control

A.3
B.1 B.2 B.3
B.6
C.5 ET.2
ET.3
ET.7 CYP.2
CYP.3
CYP.5 AP.8
AP.9
AP.11
AP.12
AP.13
AP.21 AP.22 AP.23
AP.26
Types of feedback seen

![Bar chart showing the number of practice steps for different types of feedback seen in the intervention and control groups. The x-axis represents the types of feedback, including Grammar 1, Grammar 2, Grammar 3, Other Gram., Orthography, Meaning, Default, and Correct. The y-axis represents the number of practice steps. The groups are shown with red bars for the intervention and blue bars for the control. The chart shows a significant difference in the number of practice steps seen between the two groups, with the intervention group having a higher number in most categories.]
Log data: Analyzing task interactions
Which tasks support learning for this population?

Final version fully correct?
- yes
- no

Theme 1 C, SubTask 3 [short answers]

268 Abgaben (170 vollständig korrekt, 98 fehlerhaft)
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


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
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.

SA: Der Geruch von Fisch und Schiffsdiesel beim Hafen. The smell of fish and fuel at the port.
Alignment restricted to focused constituent

Q: ‘Which sport does Isabel do?’
TA: ‘She likes to go [jogging]_F.’
SA: ‘[Jogging]_F is fun for her.’
<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>With Focus</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREG-1032</td>
<td>85.9%</td>
<td>88.6%</td>
<td>+2.7%</td>
</tr>
<tr>
<td>CREG-2155</td>
<td>82.1%</td>
<td>85.1%</td>
<td>+3.0%</td>
</tr>
<tr>
<td>Overall</td>
<td>83.2%</td>
<td>85.6%</td>
<td>+2.4%</td>
</tr>
</tbody>
</table>
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

![Graph showing the relationship between challenge (input - baseline writing) and improvement (continuation writing - baseline writing). The graph includes a line of best fit with the equation $r^2 = 2\%$.](image-url)
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

⇒ support selection of individually challenging texts
  - Syntactic Benchmark (SyB) (Chen & Meurers 2017a, 2019)
SyB: Syntactic Benchmark (http://complexityweb.org)
Introduction

Linguistic complexity: data-driven analysis
- Definition & empirical basis
- Analyzing development in large-scale data
- Task effects on complexity measures of development
- Manipulating task complexity
- Interpreting learner language

Task-driven analysis
- The FeedBook
- FeedBook characteristics
- Testing effectiveness in a randomized field trial
- Results of the field trial

Conclusions

References

Appendix
- Short answer assessment
- Complexity alignment
Introduction

Linguistic complexity: data-driven analysis
- Definition & empirical basis
- Analyzing development in large-scale data
- Task effects on complexity measures of development
- Manipulating task complexity
- Interpreting learner language

Task-driven analysis
- The FeedBook
- FeedBook characteristics
- Testing effectiveness in a randomized field trial
- Results of the field trial

Conclusions

References

Appendix
- Short answer assessment
- Complexity alignment
Introduction

Linguistic complexity: data-driven analysis
- Definition & empirical basis
- Analyzing development in large-scale data
- Task effects on complexity measures of development
- Manipulating task complexity
- Interpreting learner language

Task-driven analysis
- The FeedBook
- FeedBook characteristics
- Testing effectiveness in a randomized field trial
- Results of the field trial

Conclusions

References

Appendix
- Short answer assessment
- Complexity alignment
Introduction

Linguistic complexity: data-driven analysis
- Definition & empirical basis
- Analyzing development in large-scale data
- Task effects on complexity measures of development
- Manipulating task complexity
- Interpreting learner language

Task-driven analysis
- The FeedBook
- FeedBook characteristics
- Testing effectiveness in a randomized field trial
- Results of the field trial

Conclusions

References

Appendix
- Short answer assessment
- Complexity alignment
Introduction

Linguistic complexity: data-driven analysis
- Definition & empirical basis
- Analyzing development in large-scale data
- Task effects on complexity measures of development
- Manipulating task complexity
- Interpreting learner language

Task-driven analysis
- The FeedBook
- FeedBook characteristics
- Testing effectiveness in a randomized field trial
- Results of the field trial

Conclusions

References

Appendix
- Short answer assessment
- Complexity alignment