An Integrative Approach to Linguistic Complexity Analysis for German

Zarah Weiss

based on work done in collaboration with Detmar Meurers

Department of Linguistics & LEAD Research Network
University of Tübingen

CRC 1412 Register Colloquium, Online, June 7th, 2021
What is linguistic complexity?

- Complexity is a very productive construct in linguistics
- Application domains beyond research on language typology and second language acquisition (SLA)
  - Increasing interest in empirical education research
- Dissertation project on broad linguistic modeling of German using concept of linguistic complexity
  - Focus on automatic assessment with NLP methods
- Two main strands of research
  1. Foundational complexity research rooted in SLA and CL
  2. Application to data from ecologically valid learning contexts
An Integrative Approach to Linguistic Complexity Analysis

Zarah Weiss

Introduction

Linguistic Modeling
- System Description
- Linguistic Complexity
- Language Use
- Human Processing
- Multilingual CTAP

Foundational Complexity Research
- Proficiency Assessment
- Readability Assessment
- Linking input and output
- Application to ecologically valid education contexts

Task effects on CAF
- Predicting Tasks
- Task-Effects Across Complexity Domains

Outlook

References

Appendix

Complexity, Accuracy, and Fluency

- **CAF triad describes dimensions of language performance** (Housen et al. 2012)
  - Complexity = elaborateness and variation of language
  - Accuracy = native-like and error-free language use
  - Fluency = native-like production speed and smoothness

- **Assessed to characterize L2 proficiency and development**
  - Focus on English, but also some work on other languages (e.g., De Clercq & Housen 2019; Brezina & Pallotti 2019)

- **Recent criticism on reductionist approach to complexity** (Housen et al. 2019; Pallotti 2009, 2015)
  - Focus in lexical and syntactic complexity
  - Increasing interest in morphology and phraseology (De Clercq & Housen 2019; Brezina & Pallotti 2019; Paquot 2019)

- **Adequacy** suggested as 4th CAF dimension (e.g. Pallotti 2009)
Task Effects on Complexity, Accuracy, Fluency

- Task effects have been shown to influence CAF (Yoon & Polio 2016; Yang et al. 2015; Yoon 2017)
  - L2 corpora rarely control for comparable task conditions (Tracy-Ventura & Myles 2015; Yoon & Polio 2016)
- Debated which task factors contribute to task effects
  - Functional task factors (e.g., genre, topic)
  - Cognitive task factors (e.g., shared context, code complexity)
- Effect of cognitive task factors on CAF trade-offs debated
  - Limited Attentional Capacity Model (Skehan 1996)
  - Cognition Hypothesis (Robinson 2001)
Focus of today’s talk

1. Broad modeling of linguistic complexity across linguistic domains (syntax, lexicon, morphology, and discourse)
   ▶ Integration of psycho-linguistic measures of language use
   ▶ Integration of measures of human processing

2. Model German L2 proficiency and readability and link linguistic characteristics of reading materials and writing

3. Study role of tasks in L2 writing across linguistic domains
Automatic Linguistic Modeling

▶ Extract 489 features of German linguistic complexity, language use, human processing, and discourse cohesion
  ▶ Linguistic domains of syntax, lexicon, and morphology
▶ It has been successfully used for the assessment of, e.g.,
  ▶ Text readability (Weiss & Meurers 2018; Weiss et al. 2021)
  ▶ L2 proficiency (Weiss & Meurers 2019b, 2021)
  ▶ Academic language development (Weiss & Meurers 2019a)
  ▶ Teachers’ grading behavior (Weiss et al. 2019)
▶ Features in multilingual Common Text Analysis Platform\(^1\) (Chen & Meurers 2016; Weiss et al. 2021)
  ▶ Intersection of 200+ features for EN, DE, FR, ES
  ▶ Working on Dutch, Italian, Arabic, Portuguese, Turkish

\(^1\)http://ctapweb.com
Syntactic & Lexical Complexity

Syntactic complexity

- For clausal complexity, we use established subordination and coordination ratios (Wolfe-Quintero et al. 1998)

- For phrasal complexity, we measure phrase modification and modifier coverage (focus: NP complexity, verb cluster)

- We measure grammatical constructs of academic language such as deagentivation

Lexical complexity

- We extract traditional measures of lexical diversity and lexical variation (Lu 2012; McCarthy & Jarvis 2010)

- We subsume some word-net based measures of semantic relatedness in this category
Morphological Complexity & Discourse Measures

Morphological complexity

- We assess the elaboration of compounding and derivation
- Ratios tracking the percentage of certain inflection markers
- Morphological Complexity Index for morphological variation (Brezina & Pallotti 2019)

Discourse Measures

- We assess the use of overt cohesion markers such as certain connectives (Graesser et al. 2004; Crossley et al. 2016)
- We measure implicit cohesion generated through word overlap and grammatical transitions (McNamara et al. 2014; Barzilay & Lapata 2008; Todirascu et al. 2013)
Language Use

Why language use?

- Language use measures are motivated by human behavior rather than the theoretical linguistic system

- Examples are Age of Acquisition measures but also frequency measures (traditionally lexical complexity)

Measures in our system

- Average word and lemma frequencies and word familiarity across data bases (Brysbaert et al. 2011b; Heister et al. 2011)

- Approximations of Age of Active use based on a corpus of children’s texts (Lavalley et al. 2015)
Human Processing Costs

The Dependency Locality Theory

- Human processing costs are estimated by increased processing times caused by surprisal or cognitive load
- Dependency locality theory (Gibson 2000) prominent theory of sentence processing
  - Storage of incomplete discourse structures and discourse referents consumes resources
  - Distant discourse referents cause higher processing costs

Measures in our system

- We measure maximal and average DLT integration costs
- We include weighting configurations suggested by Shain et al. (2016) for verbs, coordination, and modifiers
An Integrative Approach to Linguistic Complexity Analysis

Zarah Weiss

Introduction

Linguistic Modeling

System Description
Linguistic Complexity
Language Use
Human Processing

Multilingual CTAP

Foundational Complexity Research

Proficiency Assessment
Readability Assessment
Linking input and output
Application to ecologically valid education contexts

Task effects on CAF

Predicting Tasks
Task-Effects Across Complexity Domains

Outlook

References

Appendix
An Integrative Approach to Linguistic Complexity Analysis

Zarah Weiss

Introduction

Linguistic Modeling

System Description

Linguistic Complexity

Language Use

Human Processing

Multilingual CTAP

Foundational Complexity Research

Proficiency Assessment

Readability Assessment

Linking input and output

Application to ecologically valid education contexts

Task effects on CAF

Predicting Tasks

Task-Effects Across Complexity Domains

Outlook

References

Appendix

CTAP dashboard (www.ctapweb.com)
CTAP feature manager (www.ctapweb.com)
An Integrative Approach to Linguistic Complexity Analysis

Zarah Weiss

Introduction

CTAP analysis manager (www.ctapweb.com)
An Integrative Approach to Linguistic Complexity Analysis

Zarah Weiss

Introduction

Linguistic Modeling

System Description

Linguistic Complexity

Language Use

Human Processing

Multilingual CTAP

Foundational Complexity Research

Proficiency Assessment

Readability Assessment

Linking input and output

Application to ecologically valid education contexts

Task effects on CAF

Predicting Tasks

Task-Effects Across Complexity Domains

Outlook

References

Appendix
Automatic proficiency assessment

- How can learners’ language proficiency be characterized through their language productions?
  - Linked to tracking of learners’ developmental trajectories
  - Similar to automatic essay grading (Crossley 2020)

- Common application domain of complexity research in SLA (Housen et al. 2019)
  - Typically used to track written language proficiency

- Previous studies on German L2/L1 proficiency assessment
  - Written L2 proficiency (Weiss & Meurers 2019b, 2021)
  - Oral or written early academic language (Weiss & Meurers 2019a; Weiss et al. in preparation b)
  - English/German advanced academic language (Chen et al. draft)
Automatic proficiency assessment

▶ How can learners’ language proficiency be characterized through their language productions?
  ▶ Linked to tracking of learners’ developmental trajectories
  ▶ Similar to automatic essay grading (Crossley 2020)

▶ Common application domain of complexity research in SLA (Housen et al. 2019)
  ▶ Typically used to track written language proficiency

▶ Previous studies on German L2/L1 proficiency assessment
  ▶ **Written L2 proficiency** (Weiss & Meurers 2019b, 2021)
  ▶ Oral or written early academic language (Weiss & Meurers 2019a; Weiss et al. in preparation b)
  ▶ English/German advanced academic language (Chen et al. draft)
The Merlin corpus

- Cross-sectional corpus of 1,033 German L2 writings (Wisniewski et al. 2013)
  - Texts rated by two experts on CEFR scale (A1 to C2)
- Elicited in official standardized language certification tests
  - Test levels A1 to C1
- Texts available as original writing and two target hypothesis
  - Form-based target hypothesis for entire corpus
  - Meaning-based target hypothesis for sub-corpus
- Three task prompts per level ⇒ 15 different task prompt
Proficiency scores are distributed across test levels, e.g., A2 and B1 test takers’ proficiency scores range from A1 to B2.
<table>
<thead>
<tr>
<th>Level</th>
<th>N. docs</th>
<th>μ words ± SD</th>
<th>M</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>57</td>
<td>34. ± 11.5</td>
<td>35</td>
<td>6</td>
<td>63</td>
</tr>
<tr>
<td>A2</td>
<td>306</td>
<td>61. ± 23.3</td>
<td>55</td>
<td>24</td>
<td>141</td>
</tr>
<tr>
<td>B1</td>
<td>331</td>
<td>118. ± 48.3</td>
<td>107</td>
<td>28</td>
<td>293</td>
</tr>
<tr>
<td>B2</td>
<td>293</td>
<td>195. ± 47.4</td>
<td>195</td>
<td>66</td>
<td>354</td>
</tr>
<tr>
<td>C1</td>
<td>42</td>
<td>227. ± 49.9</td>
<td>223</td>
<td>150</td>
<td>372</td>
</tr>
<tr>
<td>C2</td>
<td>4</td>
<td>246. ± 51.5</td>
<td>241</td>
<td>192</td>
<td>310</td>
</tr>
</tbody>
</table>

- Skewed distribution with few A1 and C1/C2 texts
- Text length systematically increases with higher proficiency
- From level B1 on most texts contain more than 100 words
Merlin corpus statistics

<table>
<thead>
<tr>
<th></th>
<th>N. docs</th>
<th>(\mu) words (\pm SD)</th>
<th>M</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>57</td>
<td>34. (\pm 11.5)</td>
<td>35</td>
<td>6</td>
<td>63</td>
</tr>
<tr>
<td>A2</td>
<td>306</td>
<td>61. (\pm 23.3)</td>
<td>55</td>
<td>24</td>
<td>141</td>
</tr>
<tr>
<td>B1</td>
<td>331</td>
<td>118. (\pm 48.3)</td>
<td>107</td>
<td>28</td>
<td>293</td>
</tr>
<tr>
<td>B2</td>
<td>293</td>
<td>195. (\pm 47.4)</td>
<td>195</td>
<td>66</td>
<td>354</td>
</tr>
<tr>
<td>C1</td>
<td>42</td>
<td>227. (\pm 49.9)</td>
<td>223</td>
<td>150</td>
<td>372</td>
</tr>
<tr>
<td>C2</td>
<td>4</td>
<td>246. (\pm 51.5)</td>
<td>241</td>
<td>192</td>
<td>310</td>
</tr>
</tbody>
</table>

▶ Skewed distribution with few A1 and C1/C2 texts

▶ Text length systematically increases with higher proficiency

▶ From level B1 on most texts contain more than 100 words
Merlin corpus statistics

<table>
<thead>
<tr>
<th>Level</th>
<th>N. docs</th>
<th>$\mu$ words $\pm$ SD</th>
<th>M</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>57</td>
<td>34. $\pm$ 11.5</td>
<td>35</td>
<td>6</td>
<td>63</td>
</tr>
<tr>
<td>A2</td>
<td>306</td>
<td>61. $\pm$ 23.3</td>
<td>55</td>
<td>24</td>
<td>141</td>
</tr>
<tr>
<td>B1</td>
<td>331</td>
<td>118. $\pm$ 48.3</td>
<td>107</td>
<td>28</td>
<td>293</td>
</tr>
<tr>
<td>B2</td>
<td>293</td>
<td>195. $\pm$ 47.4</td>
<td>195</td>
<td>66</td>
<td>354</td>
</tr>
<tr>
<td>C1</td>
<td>42</td>
<td>227. $\pm$ 49.9</td>
<td>223</td>
<td>150</td>
<td>372</td>
</tr>
<tr>
<td>C2</td>
<td>4</td>
<td>246. $\pm$ 51.5</td>
<td>241</td>
<td>192</td>
<td>310</td>
</tr>
</tbody>
</table>

- Skewed distribution with few A1 and C1/C2 texts
- Text length systematically increases with higher proficiency
- From level B1 on most texts contain more than 100 words
## Merlin corpus statistics

<table>
<thead>
<tr>
<th></th>
<th>N. docs</th>
<th>$\mu$ words ± SD</th>
<th>M</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>57</td>
<td>34. ± 11.5</td>
<td>35</td>
<td>6</td>
<td>63</td>
</tr>
<tr>
<td>A2</td>
<td>306</td>
<td>61. ± 23.3</td>
<td>55</td>
<td>24</td>
<td>141</td>
</tr>
<tr>
<td>B1</td>
<td>331</td>
<td>118. ± 48.3</td>
<td>107</td>
<td>28</td>
<td>293</td>
</tr>
<tr>
<td>B2</td>
<td>293</td>
<td>195. ± 47.4</td>
<td>195</td>
<td>66</td>
<td>354</td>
</tr>
<tr>
<td>C1</td>
<td>42</td>
<td>227. ± 49.9</td>
<td>223</td>
<td>150</td>
<td>372</td>
</tr>
<tr>
<td>C2</td>
<td>4</td>
<td>246. ± 51.5</td>
<td>241</td>
<td>192</td>
<td>310</td>
</tr>
</tbody>
</table>

- Skewed distribution with few A1 and C1/C2 texts
- Text length systematically increases with higher proficiency
- From level B1 on most texts contain more than 100 words
Merlin corpus statistics

<table>
<thead>
<tr>
<th></th>
<th>N. docs</th>
<th>μ words ± SD</th>
<th>M</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>57</td>
<td>34. ± 11.5</td>
<td>35</td>
<td>6</td>
<td>63</td>
</tr>
<tr>
<td>A2</td>
<td>306</td>
<td>61. ± 23.3</td>
<td>55</td>
<td>24</td>
<td>141</td>
</tr>
<tr>
<td>B1</td>
<td>331</td>
<td>118. ± 48.3</td>
<td>107</td>
<td>28</td>
<td>293</td>
</tr>
<tr>
<td>B2</td>
<td>293</td>
<td>195. ± 47.4</td>
<td>195</td>
<td>66</td>
<td>354</td>
</tr>
<tr>
<td>C1</td>
<td>42</td>
<td>227. ± 49.9</td>
<td>223</td>
<td>150</td>
<td>372</td>
</tr>
<tr>
<td>C2</td>
<td>4</td>
<td>246. ± 51.5</td>
<td>241</td>
<td>192</td>
<td>310</td>
</tr>
</tbody>
</table>

- Skewed distribution with few A1 and C1/C2 texts
- Text length systematically increases with higher proficiency
- From level B1 on most texts contain more than 100 words
Machine learning experiment on Merlin

- Predict CEFR levels A1 to C1/C2 (Weiss & Meurers 2019b)
- Extract 489 features from form-based target hypothesis
- Training and testing with 10-fold cross-validation
- Compare (ordinal) random forest, SVM (polynomial/radial)
  - Ordinal random forest performs best
- Accuracy = 70.0% in 5-way classification
  - Majority baseline: 32.0%
Automatic readability assessment

- Readability assessment identifies suitable texts given
  - the language proficiency of the target audience
  - the reading purpose (e.g., learning, information retrieval)

- Complexity features successfully predict readability
  (Vajjala & Meurers 2012; Chen & Meurers 2017b; Crossley et al. 2019)

- Complexity-based readability research on German
  - Distinction of media language for adults and children
    (Weiss & Meurers 2018)
  - Readability research on schoolbooks
    (Kühberger et al. 2019; Berendes et al. 2018)
  - First multi-level readability classification for L2 materials
    (Weiss et al. 2021)

- Linking input and output
  Application to ecologically valid education contexts

- Task effects on CAF
  Predicting Tasks
  Task-Effects Across Complexity Domains

- Outlook

- References

- Appendix
Automatic readability assessment

- Readability assessment identifies suitable texts given
  - the language proficiency of the target audience
  - the reading purpose (e.g., learning, information retrieval)

- Complexity features successfully predict readability
  (Vajjala & Meurers 2012; Chen & Meurers 2017b; Crossley et al. 2019)

- Complexity-based readability research on German
  - Distinction of media language for adults and children
    (Weiss & Meurers 2018)
  - Readability research on schoolbooks
    (Kühberger et al. 2019; Berendes et al. 2018)
  - First multi-level readability classification for L2 materials
    (Weiss et al. 2021)
Spotlight-DE data set

- **1,446 articles** from monthly language learning magazine *Deutsch perfekt* (www.deutsch-perfekt.de)
  - Published in 45 issues from 01/2018 to 12/2019
  - Extract article from type-set PDFs using OCR
- **Leveled reading materials** aligned with CEFR levels
  - Beginning (A2), medium (B1/B2), advanced (C1)
- Part of multi-lingual data set for English and German
  - Being extended to French, Spanish, Italian
  ⇒ All data from magazines by Spotlight publisher
### Spotlight-DE text statistics

<table>
<thead>
<tr>
<th></th>
<th>N. docs</th>
<th>μ words ± SD</th>
<th>M</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>763</td>
<td>236±235</td>
<td>137</td>
<td>60</td>
<td>1.469</td>
</tr>
<tr>
<td>Medium</td>
<td>509</td>
<td>665±769</td>
<td>448</td>
<td>72</td>
<td>5.605</td>
</tr>
<tr>
<td>Advanced</td>
<td>174</td>
<td>892±537</td>
<td>524</td>
<td>91</td>
<td>4.161</td>
</tr>
</tbody>
</table>

- Contains more easy and medium than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts
Spotlight-DE text statistics

<table>
<thead>
<tr>
<th>Level</th>
<th>N. docs</th>
<th>$\mu$ words ± SD</th>
<th>M</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>763</td>
<td>236±235</td>
<td>137</td>
<td>60</td>
<td>1.469</td>
</tr>
<tr>
<td>Medium</td>
<td>509</td>
<td>665±769</td>
<td>448</td>
<td>72</td>
<td>5.605</td>
</tr>
<tr>
<td>Advanced</td>
<td>174</td>
<td>892±537</td>
<td>524</td>
<td>91</td>
<td>4.161</td>
</tr>
</tbody>
</table>

- Contains more easy and medium than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts
### Spotlight-DE text statistics

<table>
<thead>
<tr>
<th></th>
<th>N. docs</th>
<th>$\mu$ words ± SD</th>
<th>M</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>763</td>
<td>236±235</td>
<td>137</td>
<td>60</td>
<td>1.469</td>
</tr>
<tr>
<td>Medium</td>
<td>509</td>
<td>665±769</td>
<td>448</td>
<td>72</td>
<td>5.605</td>
</tr>
<tr>
<td>Advanced</td>
<td>174</td>
<td>892±537</td>
<td>524</td>
<td>91</td>
<td>4.161</td>
</tr>
</tbody>
</table>

- Contains more easy and medium than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts
Spotlight-DE text statistics

<table>
<thead>
<tr>
<th>N. docs</th>
<th>$\mu$ words $\pm SD$</th>
<th>M</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>763</td>
<td>236±235</td>
<td>137</td>
<td>60 1.469</td>
</tr>
<tr>
<td>Medium</td>
<td>509</td>
<td>665±769</td>
<td>448</td>
<td>72 5.605</td>
</tr>
<tr>
<td>Advanced</td>
<td>174</td>
<td>892±537</td>
<td>524</td>
<td>91 4.161</td>
</tr>
</tbody>
</table>

- Contains more easy and medium than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts
Machine learning experiment on Spotlight-DE

- Automatic prediction of reading levels A2, B1/B2, and C1 (Weiss et al. 2021)
- Extract 489 features from plain text articles
- Training and testing with 10-fold cross-validation
- Compare (ordinal) random forest, SVM (polynomial/radial)
  - Ordinal random forest performs best
- **Accuracy = 89.0%** in 3-way classification
  - Majority baseline: 52.8%
Appropriate input fosters language learning

- Language input fosters (second) language learning
  - Challenging input at i+1 (Krashen 1981)
  - Zone of Proximal Development (ZPD; Lantolf et al. 2015)

- Identification of appropriate reading materials requires
  1. Assessment of proficiency of language learner
  2. Assessment of reading level of reading materials
  3. Match of proficiency and reading level
  ⇒ Reading level not necessarily at i+1 for every learner (Chen & Meurers 2017a)

- How do German L2 proficiency and reading levels match together from linguistic perspective?
Linking L2 language output and L2 reading input

- 489 features successfully discriminate CEFR levels on reading texts and written learner productions
- Compare expression of relevant features on both data sets
  - Identify informative features with information gain ranking
  - Select 422 features that are informative for both data sets
- Zoom in on features across linguistic domains to compare feature values in reading and writing data
  - Syntax, lexicon, morphology, text cohesion, language use
An Integrative Approach to Linguistic Complexity Analysis
Zarah Weiss

Introduction
Linguistic Modeling
System Description
Linguistic Complexity
Language Use
Human Processing
Multilingual CTAP

Foundational Complexity Research
Proficiency Assessment
Readability Assessment

Linking input and output
Application to ecologically valid education contexts

Task effects on CAF
Predicting Tasks
Task-Effects Across Complexity Domains

Outlook
References
Appendix

Zooming in on syntactic complexity

- Systematic increase syntactic elaborateness in both
- Values for reading texts are systematically lower

⇒ No syntactic challenge in reading texts

Mean Length of Noun Phrase
Dependent clause ratio

Corpus read write

A1 A2 B1 B1/B2 B2 C1

CEFR Score

Value

0.1
0.2
0.3
0.4

0.1
0.2
0.3
0.4

A1 A2 B1 B1/B2 B2 C1

CEFR Score

Value
Zooming in on lexical complexity

- Systematic increase in lexical variation for both
- MTLD and corrected verb variation higher for reading texts
  - Challenge for learners in lexical domain
Zooming in on morphological complexity

- Systematic increase in morphological variation for both
- MCI of reading texts challenging for A2 and B1 learners
- No challenge for intermediate/advanced learners
Reading texts systematically more use of connectives and higher implicit cohesion

Challenge with respect to text cohesion
Zooming in on language use

- Reading texts use more rare words and more words with higher familiarity at constant rate
- Systematic increase in writing data with intermediate to advanced learners showing same level of familiarity

⇒ Challenge with respect to language use
Discussion

- Reading input lies systematically above language production in all domains but syntax
- Is this challenge indeed in the ZPD or above?
  - Study by Chen & Meurers (2017a) for individual features
- Is there a trade-off of acceptable challenges between different linguistic domains?
- Research program needed on interaction between proficiency and input to foster learning
Moving beyond foundational research

- Linguistic complexity as valuable concept for understanding of learning processes and outcomes
- More research needed to understand implications for real-life teaching and learning
- Studies on data from ecologically valid contexts needed
- Interdisciplinary work to connect insights from complexity research with empirical education research
CA(F) in authentic education contexts

- Study linguistic complexity in oral classroom interactions
  - Student utterances mirror written language proficiency (Weiss et al. in preparation b)
  - Complexity of teacher utterances differs across school types (Weiss et al. in preparation a)
  - Teachers’ adaptation skills in L2 classrooms (COLD project)

- Complexity & genre impact **text item difficulty** in reading comprehension tests in grade 3 & 4 (Ludewig et al. in preparation)

- It links to complexity of historical thinking processes
  - ...displayed student writing (Bertram et al. in preparation)
  - ...required by tasks history books (Kühberger et al. 2019)

- Task-appropriate complexity and accuracy influence teachers’ grading behavior in Abitur data (Weiss et al. 2019)

---

2 www.die-bonn.de/cold
Narrowing in on task effects

- Describe interaction of task and complexity
  - How do task effects differ across linguistic domains?

- Use Merlin corpus for investigation of task effects
  1. Predict 15 tasks to gauge strength of overall effect
  2. Zoom in on task effects across linguistic domains

- Focus on task rather than individual task factors
  - Annotated 15 task factors for task prompts\(^3\) (Weiss 2017)
  - Used approach by Alexopoulou et al. (2017)
  \(\Rightarrow\) Nearly all factors fully confounded with test level

---

\(^3\)Cognitive factors: code and cognitive complexity, shared context, reasoning demands, referenced elements, perspective; functional factors: genre, audience, task theme and type
Writing formality confounds with test level

Proficiency Scores and Formality in Merlin

- **Number of Essays**
- **Overall CEFR Score**
- **Formality**
  - formal
  - informal

Overall CEFR Score

- A1
- A2
- B1
- B2
- C1

Formality: formal and informal
Text type confounds with test level
Code complexity confounds with test level

- Code complexity depends on provision of no, few, or detailed language material in instructions.
Cognitive complexity confounds with test level

Did the task clearly outline the intended text content and structure or did learners need to structure texts while writing?
Task classification on Merlin corpus

- Automatic prediction of 15 tasks (3 per test level)
- Extract 489 features from form-based target hypothesis
- Training and testing with 10-fold cross-validation
- Compare random forest, SVM (polynomial/radial)
  - SVM with polynomial kernel performs best
- Accuracy = 89.8% in 15-way classification
  - Majority baseline: 8.0%
  - Exceeds 70.0% accuracy for 5-way proficiency classification
Zooming in on the Merlin corpus

- Subset of the Merlin corpus used for classification studies
- Focus on B1/B2 learners taking B1/B2 tests (N=337)
  - Overall six different elicitation tasks
  - This also includes, e.g., B1 learners who failed a B2 test

---

**Proficiency Scores in Merlin across Task Prompts**

- B1: Happy New Year
- B1: Happy birthday
- B1: Visiting a friend
- B2: Aupair complaint
- B2: Aupair request
- B2: Internship

**Overall CEFR Score**

<table>
<thead>
<tr>
<th>Task Prompt</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy New Year</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>Happy birthday</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>Visiting a friend</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Aupair complaint</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Aupair request</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>Internship</td>
<td>30</td>
<td>20</td>
</tr>
</tbody>
</table>

**Number of Essays**

- B1: Happy New Year: 20
- B1: Happy birthday: 10
- B1: Visiting a friend: 0
- B2: Aupair complaint: 50
- B2: Aupair request: 40
- B2: Internship: 30
B2 texts use more elaborate subordination than B1 texts
- This holds across measures of clausal complexity

We see some some task-specific variation at B1 and B2
- Specific tasks elicit more or less clausal elaboration
- This holds across measures of clausal complexity
⇒ Verb phrase complexity is influenced by test level but also shows differences between B1 and B2 scores
   ▶ No clear proficiency differences for NP or PP complexity
Phrasal Complexity

Verb phrase complexity is influenced by test level but also shows differences between B1 and B2 scores.

- No clear proficiency differences for NP or PP complexity.
Verb phrase complexity is influenced by test level but also shows differences between B1 and B2 scores.

- No clear proficiency differences for NP or PP complexity
B2 texts have a higher lexical variation than B1 texts

- Some tasks have a very strong effect on lexical variation
  - Lexical diversity (MTLD) shows no significant differences
  - Semantic relatedness is highly sensitive to task differences
The use of derived forms is highly task dependent but for most tasks, B2 texts contain more derivation than B1 texts.

Most stable morphological complexity feature we found:
- Use of composition more task than proficiency dependent
- Inflection features show even stronger task dependence
Discourse Complexity

⇒ B2 test takers write more implicitly cohesive texts
  ▶ Limited within task difference between B1 and B2 texts

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing
  ▶ B2 tests were more often taken by B2 than B1 learners
An Integrative Approach to Linguistic Complexity Analysis

Zarah Weiss

Introduction

Linguistic Modeling
System Description
Linguistic Complexity
Language Use
Human Processing
Multilingual CTAP

Foundational Complexity Research
Proficiency Assessment
Readability Assessment
Linking input and output
Application to ecologically valid education contexts

Task effects on CAF
Predicting Tasks
Task-Effects Across Complexity Domains

Outlook
References
Appendix

Discourse Complexity

Global Stem Overlap per Sentence

⇒ B2 test takers write more implicitly cohesive texts
   ▶ Limited within task difference between B1 and B2 texts

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing
   ▶ B2 tests were more often taken by B2 than B1 learners
B2 test takers write more implicitly cohesive texts
  ▶ Limited within task difference between B1 and B2 texts

Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing
  ▶ B2 tests were more often taken by B2 than B1 learners
Tasks for B2 testing elicit more rare vocabulary

- Within tasks, texts by B1 and B2 learners do not differ
- This pattern holds for all frequency features we measured

Again, overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing
Tasks for B2 testing elicit more rare vocabulary
  - Within tasks, texts by B1 and B2 learners do not differ
  - This pattern holds for all frequency features we measured

Again, overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing
Tasks for B2 testing elicit more rare vocabulary

- Within tasks, texts by B1 and B2 learners do not differ
- This pattern holds for all frequency features we measured

Again, overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing
Human Processing Costs

⇒ Integration costs are higher for texts with B2 scores
⇒ Within proficiency levels, they are stable across tasks
Discussion

- Comparison showed that some presumed proficiency differences can be explained by task differences
  - Morphological complexity
  - Discourse complexity
  - Language use

- Some measures of lexical and phrasal complexity characterize proficiency differences despite task sensitivity

- Two complexity domains were systematically stable
  - Human processing features
  - Clausal complexity features

- Which (combination of) task factors causes these effects?
  - More research and data needed
Conclusion

- System for broad linguistic modeling of linguistic complexity, language use, processing, and discourse
  - Applicable to German, English, French, Spanish
  - Being extended to Italian, Turkish, Arabic, Portuguese

- Characterized differences in proficiency and reading differences to compare L2 input and output
  - Reading input lies systematically above language production in all domains but syntax
  - Is this challenge indeed in the ZPD or above?

- Showed influence of tasks on complexity in L2 writing
  - Demonstrated remarkable role of tasks on complexity
  - Morphology, language use & discourse task dependent
  - Processing and clausal features seem more stable
  - Role of individual task factors remains unclear
Outlook

- Specify interaction between proficiency & appropriate input
  - Is there a trade-off of acceptable challenges between different linguistic domains?
  - Study link in spoken language, e.g., classroom interactions (see COLD project, www.die-bonn.de/cold)

- Interaction of task factors, proficiency, and linguistic domains
  - How does proficiency mediate task effects on CAF?
  - Extension across languages to test generalizability of findings
  - Separating task factors to understand cause of task effects
Outlook

- Specify interaction between proficiency & appropriate input
  - Is there a trade-off of acceptable challenges between different linguistic domains?
  - Study link in spoken language, e.g., classroom interactions (see COLD project, www.die-bonn.de/cold)

- Interaction of task factors, proficiency, and linguistic domains
  - How does proficiency mediate task effects on CAF?
  - Extension across languages to test generalizability of findings
  - Separating task factors to understand cause of task effects
References


Brysbaert, M., M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte & A. Böhl (2011a). The word frequency effect: A review of recent developments and implications...


An Integrative Approach to Linguistic Complexity Analysis
Zarah Weiss

Introduction

Linguistic Modeling
System Description
Linguistic Complexity
Language Use
Human Processing
Multilingual CTAP

Foundational Complexity Research
Proficiency Assessment
Readability Assessment
Linking input and output
Application to ecologically valid education contexts

Task effects on CAF
Predicting Tasks
Task-Effects Across Complexity Domains

Outlook

References

Appendix


An Integrative Approach to Linguistic Complexity Analysis
Zarah Weiss

Introduction

Linguistic Modeling
System Description
Linguistic Complexity
Language Use
Human Processing
Multilingual CTAP

Foundational Complexity Research
Proficiency Assessment
Readability Assessment
Linking input and output
Application to ecologically valid education contexts

Task effects on CAF
Predicting Tasks
Task-Effects Across Complexity Domains

Outlook

References

Appendix


An Integrative Approach to Linguistic Complexity Analysis

Zarah Weiss

Introduction

Linguistic Modeling

System Description

Linguistic Complexity

Language Use

Human Processing

Multilingual CTAP

Foundational Complexity Research

Proficiency Assessment

Readability Assessment

Linking input and output

Application to ecologically valid education contexts

Task effects on CAF

Predicting Tasks

Task-Effects Across Complexity Domains

Outlook

References

Appendix


An Integrative Approach to Linguistic Complexity Analysis
Zarah Weiss

Introduction

Linguistic Modeling
System Description
Linguistic Complexity
Language Use
Human Processing
Multilingual CTAP

Foundational Complexity Research
Proficiency Assessment
Readability Assessment
Linking input and output
Application to ecologically valid education contexts

Task effects on CAF
Predicting Tasks
Task-Effects Across Complexity Domains

Outlook

References

Appendix


How do we define complexity?

Figure from Bulté & Housen (2014)
Taxonomy of Complexity

Figure from Housen et al. (2012)
Taxonomy of Complexity

Figure from Housen et al. (2012)
An Integrative Approach to Linguistic Complexity Analysis
Zarah Weiss

Introduction

Linguistic Modeling
System Description
Linguistic Complexity
Language Use
Human Processing
Multilingual CTAP

Foundational Complexity Research
Proficiency Assessment
Readability Assessment
Linking input and output
Application to ecologically valid education contexts

Task effects on CAF
Predicting Tasks
Task-Effects Across Complexity Domains

Outlook
References
Appendix

Taxonomy of Complexity

Figure from Housen et al. (2012)
Taxonomy of Complexity

Figure from Housen et al. (2012)
An Integrative Approach to Linguistic Complexity Analysis

Zarah Weiss

Introduction
Linguistic Modeling
System Description
Linguistic Complexity
Language Use
Human Processing
Multilingual CTAP

Foundational Complexity Research
Proficiency Assessment
Readability Assessment
Linking input and output
Application to ecologically valid education contexts

Task effects on CAF
Predicting Tasks
Task-Effects Across Complexity Domains

Outlook
References
Appendix

Taxonomy of Complexity

Figure from Housen et al. (2012)
An Integrative Approach to Linguistic Complexity Analysis
Zarah Weiss

Introduction

Linguistic Modeling
- System Description
- Linguistic Complexity
- Language Use
- Human Processing
- Multilingual CTAP

Foundational Complexity Research
- Proficiency Assessment
- Readability Assessment
- Linking input and output
- Application to ecologically valid education contexts

Task effects on CAF
- Predicting Tasks
- Task-Effects Across Complexity Domains

Outlook

References

Appendix

Taxonomy of Complexity

Figure from Housen et al. (2012)
Taxonomy of Complexity

Figure from Housen et al. (2012)
Broad multi-lingual complexity modeling

1. Shared NLP pipeline to obtain
   - Sentences, words, POS tags, constituency/dependency parses with Stanford CoreNLP (Manning et al. 2014)
   - Lemmas with Mate tools (Bohnet & Nivre 2012)
   - Stems with OpenNLP Snowball stemmer

2. Identify linguistic constructions with extraction rules
   - Language-specific rules for syllables, POS, constituents
   - Language-independent rules for all other
   - Frequency features based on Subtlex-US and Subtlex-DE (Brysbaert et al. 2011a,b)

3. Fully language-independent calculation of feature ratios
An Integrative Approach to Linguistic Complexity Analysis

Proficiency split by tasks in Merlin

Proficiency Scores in Merlin across Task Prompts

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1: Housing office</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2: Pet sitting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1: Happy birthday</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1: Happy New Year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1: Visiting a friend</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2: Aupair complaint</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2: Aupair request</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2: Internship</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1: Housing situation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1: Learning German</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1: Traditions &amp; Assimilation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Overall CEFR Score

Number of Essays

Proficiency Scores in Merlin across Task Prompts

<table>
<thead>
<tr>
<th>Task Prompt</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1: Birth congratulations</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1: Going swimming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1: Request help</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2: Housing office</td>
<td>20</td>
<td>40</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>A2: Pet sitting</td>
<td>20</td>
<td>40</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>A2: Ticket offer</td>
<td>20</td>
<td>40</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>B1: Happy birthday</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1: Happy New Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1: Visiting a friend</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2: Aupair complaint</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2: Aupair request</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2: Internship</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1: Housing situation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1: Learning German</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1: Traditions &amp; Assimilation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Outlook

References

Appendix
Falko Georgetown corpus

- Corpus of 209 texts written by 123 German L2 learners (Siemen et al. 2006)
- Elicited in curricular writing courses at the Georgetown University ranging from medium (1) to advanced (4)
  - Take-home final exams written by students
  - Task prompt either curriculum dependent or reference task
  - Several students attended consecutive courses
- No additional external validation of proficiency available
Students systematically use more dependent clauses per clause with increasing course levels

- This holds across measures of clausal complexity

Relative stability within a course level

- More pronounced differences for some features
Phrasal Complexity

⇒ VP length increases across course levels relatively stable
   ▶ PP and NP complexity features are more task dependent and show a less clear development on the reference task
More lexically varied book reviews at higher course levels

Curriculum tasks elicit more lexically varied writing
  - Lexical diversity shows a comparable pattern
  - Semantic relatedness is only sensitive to task differences
Morphological Complexity

⇒ At higher course levels, learners use more derivation but task differences are also pronounced.
⇒ Differences for inflection and composition features can all be attributed to task differences.
Discourse Complexity

⇒ Only curriculum tasks show a systematic increase in implicit cohesion while book reviews remain stable
  ▶ Similar pattern across measures of implicit cohesion
The vocabulary frequency of book reviews is relatively stable across course levels (except level 4)

- At level 4 significant frequency decrease for some features

Curriculum-tasks strongly differ vocabulary frequency from each other and the reference task
Human Processing Costs

Integration costs increase with course levels

Curriculum and reference tasks exhibit comparable values
  ▶ But they are systematically higher in curriculum-tasks
  ▶ Less pronounced difference between course levels 1 and 2