Using Measures of Linguistic Complexity to Assess German L2 Proficiency in Learner Corpora under Consideration of Task-Effects

Zarah Weiss
Eberhard Karls Universität Tübingen

Kolloquium Korpuslinguistik und Phonetik (SS17)
Humboldt Universität zu Berlin

June 7th, 2017
Table of Contents

1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
Introduction

Overview

• Complexity analysis productive approach in SLA - CL intersection
• Automatic assessment of proficiency, readability, essay scoring, etc.
• Started in 1930s with superficial measures (text length in words, word length in characters) (Frogner 1933; Thorndike 1921)
• Development towards more sophisticated easily obtainable feature sets with progress in NLP
  → Increasing number of diverse complexity measures
  → Increasing availability of text analysis systems facilitating complexity analysis
  → Lack of consistency in findings and interpretation of measures
Introduction

Criticism

• Lack of theoretical foundation of feature implementation, selection, and interpretation (Bulté and Housen 2014; Housen, Vedder, and Kuiken 2012; Pallotti 2009)

• Assumption of homogeneous learner profiles (Crossley and McNamara 2011; Jarvis et al. 2003)

• Disregard for task differences and task effects (Alexopoulou et al. 2017; Tracy-Ventura and Myles 2015)

Solutions

• Modeling of heterogeneous learner profiles; e.g. L1 backgrounds (Crossley and McNamara 2011), learning strategies (Jarvis et al. 2003)

• Modeling of task effects (Alexopoulou et al. 2017; Polio and Park 2016; Tracy-Ventura and Myles 2015)
Research Questions

1. How do measures of complexity model German L2 proficiency?
2. To which extend is this influenced by cognitive or functional task-effects?
3. Does a retrospective analysis of German learner corpora with diverse task backgrounds improve complexity-based L2 proficiency modeling?
Procedure

- Selection of 2 German learner corpora with diverse task backgrounds: Merlin (Abel et al. 2013), Falko Georgetown (Falko Georgetown Dokumentation 2007)
- Automatic extraction of 398 measures of complexity
- Extract data from close transcription of learner data
- Manual extraction of cognitive and functional task factors from task descriptions provided in corpus documentations
- Descriptive cross-corpus analysis for +100 measures
- Inferential GAM regression analysis with task interactions for data-driven feature selection
- Within-corpus classification experiment on Merlin
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
Complexity in SLA

Overview

• Assessment of language performance to measure
  • Text readability
  • L2 and L1 proficiency
  • L2 and L1 development
  • Writing performance
  • Task performance
  • ...

• Dimensions of language performance
  1. Complexity
  2. Accuracy
  3. Fluency

→ CAF
Complexity in SLA
CAF Definitions

• e.g. production rate units in time (if available); sometimes: frequency or length of sentences, t-units, utterances, clauses, and phrases

• e.g. error-free t-units, overall error count

Complexity: elaborateness, variedness, and inter-relatedness of a system (Ellis and Barkhuizen 2005; Rescher 1998)
• e.g. type token ratio, word frequency, etc. for lexical complexity
• e.g. modifiers per phrase, words per clause, clauses per t-unit, etc. for syntactic complexity
Complexity in SLA
CAF Criticism

Vagueness of definitions used in empirical studies
  • Does *sentence length* measure fluency or complexity? (Wolfe-Quintero, Inagaki, and Kim 1998)
  • What is more complex: *adäquat* or *angemessen*?

Lack of explicit norm in complexity definition
  • Is the observed complexity *adequate*? (Pallotti and Ferrari 2008; Pallotti 2015)
  • High complexity in simple tasks indicates lack of socio-linguistic awareness Ortega 2003; Pallotti 2015
Lack of theoretical underpinnings for measures

Complexity $\neq$ continuous, temporal progression

Complexity $\neq$ competence

Complexity $\neq$ difficulty

**Figure:** L2 complexity and related constructs (Bulté and Housen 2014)
Figure: Taxonomy by Housen, Vedder, and Kuiken 2012
• Complexity: general validity across groups
• Difficulty: valid for specific groups (ability dependent)
• Cf. task complexity vs. task difficulty by Robinson 2001

Figure: Taxonomy by ibid.
(1) The small, happy bear from Peru ate all the orange marmalade.

- Structure complexity: prenominal modifiers, postnominal modifiers
- System complexity: NP length, number of modifiers

**Figure:** Taxonomy by Housen, Vedder, and Kuiken 2012
(2) She carries her dogs

- 3rd P. Sg. Pres. ($s_1$) vs. plural marker ($s_2$)
- Same formal complexity
- $s_1$ functionally more complex

**Figure:** Taxonomy by Housen, Vedder, and Kuiken 2012
Taxonomy

Benefits

• Taxonomy is not exhaustive
• Helps to clarify how measures relate to complexity
• Helps to interpret findings
• Helps to assess coverage of measures included in studies
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
• In task-based language teaching CAF as dimensions of task performance
• Task factors known to influence CAF measures from research on TBLT (Polio and Park 2016; Robinson 2001; Skehan 1996)
• Task effects rarely analyzed in corpus-based studies (Tracy-Ventura and Myles 2015)
• Functional task factors vs. cognitive task factors
  → Cognitive task factors assumed to influence attentional resources dedicated to CAF
  → Functional task factors assumed to functionally require or inhibit aspects of CAF
Cognitive Task Factors

Skehan’s Limited Attentional Capacity Model

- Differentiates
  1. Code complexity
  2. Cognitive complexity (processing cost, familiarity)
  3. Communicative stress (mode, time, stakes, etc.)

→ Deplete attentional resources from CAF

Robinson’s Cognition Hypothesis

- Differentiates
  1. Task complexity
  2. Task conditions
  3. Task difficulty

→ more elements, tempo-spatial dislocation, reasoning demands direct resources to complexity and accuracy

→ more planning time, num. tasks, prior knowledge depletes resources from complexity and accuracy
Functional Task Factors

- LACM and CH make contradicting predictions about task effects on CAF
- Findings are inconsistent (Tracy-Ventura and Myles 2015; Yoon and Polio 2016)
- Functional task effects argued to be stronger (Biber, Gray, and Poonpon 2011; Yoon and Polio 2016)
- Assumed to functionally require or inhibit aspects of CAF
- Examples: discourse type, text genre, topic
- Most often comparison of argumentative, narrative, and descriptive texts
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
• Cross-sectional corpus of 1,033 German L2 writings
• Elicited in official standardized language certification tests by Abel et al. 2013
• Test levels ranging from A1 to C1
• Proficiency scores assigned by 2 expert raters ranging from A1 to C2

**Figure:** Distribution of proficiency scores grouped by test level.
## Merlin Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Test</th>
<th>$\sum$</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Going swimming</td>
<td>A1</td>
<td>56</td>
<td>8</td>
<td>45</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Apartment search</td>
<td>A1</td>
<td>77</td>
<td>11</td>
<td>50</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Child birth</td>
<td>A1</td>
<td>74</td>
<td>25</td>
<td>41</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ticket offer</td>
<td>A2</td>
<td>66</td>
<td>5</td>
<td>28</td>
<td>31</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pet sitting</td>
<td>A2</td>
<td>72</td>
<td>0</td>
<td>32</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Housing office</td>
<td>A2</td>
<td>70</td>
<td>4</td>
<td>43</td>
<td>22</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Announce visit</td>
<td>B1</td>
<td>67</td>
<td>2</td>
<td>31</td>
<td>29</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Happy birthday</td>
<td>B1</td>
<td>70</td>
<td>0</td>
<td>24</td>
<td>38</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Happy new year</td>
<td>B1</td>
<td>73</td>
<td>2</td>
<td>10</td>
<td>54</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Application</td>
<td>B2</td>
<td>69</td>
<td>0</td>
<td>1</td>
<td>22</td>
<td>42</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Work complaint</td>
<td>B2</td>
<td>70</td>
<td>0</td>
<td>1</td>
<td>20</td>
<td>47</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Information request</td>
<td>B2</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>41</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Housing situation</td>
<td>C1</td>
<td>72</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>52</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Learning German</td>
<td>C1</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>26</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>Traditions &amp; Assimilation</td>
<td>C1</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>62</td>
<td>12</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table:** Mapping of tasks to test levels, task frequency, and their distribution across overall proficiency scores (A1 to C2).
- Partially longitudinal corpus of 209 German L2 writings by 123 students
- Elicited in curricular writing courses at Georgetown University by *Falko Georgetown Dokumentation* 2007; Reznicek et al. 2007
- Course levels 1 to 4 for intermediate to advanced learners of German
- No external validation of proficiency besides course levels

**Figure:** Texts written by learners who contributed multiple writings to Falko Georgetown data plotted by semester and grouped by course level.
### Table: Frequency of tasks across course levels in the *Falko Georgetown L2* corpus.

<table>
<thead>
<tr>
<th>Task</th>
<th>∑</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write a letter</td>
<td>21</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Continue a novel</td>
<td>28</td>
<td>0</td>
<td>28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Write an article</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>Write a speech</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Book review</td>
<td>116</td>
<td>19</td>
<td>25</td>
<td>23</td>
<td>49</td>
</tr>
</tbody>
</table>
1. Introduction

2. Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3. Data
   Corpora
   Task Annotations

4. Measuring Complexity
   Automatic System
   Complexity Measures

5. Descriptive Cross-Corpus Analysis

6. Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7. Conclusion
Overview

• Annotate 15 cognitive and functional task factors
• Goal: disentangle correlation of course / test level and task
• Based on task descriptions in supplementary material *Falko Georgetown Dokumentation* 2007; Merlin project 2014a,b,c,d,e,f,g,h,i,j,k,l,m,n,o
• Follow approach by Alexopoulou et al. 2017
Operationalizations
Cognitive Factors

Code complexity: instructions provided no, few, or detailed language material to draw from

Cognitive complexity: require reasoning about writing structure vs. outline or refer to known structure

Shared context: temporal and spatial dislocation (here/there; now/then)

Reasoning demands: quantity and elaborateness of spatial reasoning, i.e. referencing a location without extra-linguistic support, and reasoning about other people’s intentions, beliefs, desires, or relationships

Referenced elements: number of discourse referents minimally required in solution

Perspective requires perspective of i) self, ii) someone else; iii) multiple other people.
Operationalizations

Functional Factors

Genre  text category; cf. task descriptions
Audience recipient; cf. task descriptions (partially grouped)
Formality tone; cf. task descriptions or inferred from genre and audience
Task theme general topic; professional/occupational interests, public social affairs, small talk, or (by extension) goal-oriented personal matters (demand)
Task type determined by a combination of functional needs and genre; argumentative, narrative, descriptive/expositional, and instructional
Table: Properties and task factors annotated for *Merlin* tasks.
### Tasks in Falko Georgetown

<table>
<thead>
<tr>
<th>Task</th>
<th>Test Level</th>
<th>Audience</th>
<th>Genre</th>
<th>Formality</th>
<th>Theme</th>
<th>Task Type</th>
<th>Code Complexity</th>
<th>Cognitive Complexity</th>
<th>Shared Context</th>
<th>Reasoning</th>
<th>Referenced Elements</th>
<th>Perspective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write a letter</td>
<td>1</td>
<td>Friend</td>
<td>Letter</td>
<td>Informal</td>
<td>Small talk</td>
<td>Instructional</td>
<td>High</td>
<td>Low</td>
<td>T &amp; T</td>
<td>Low</td>
<td>Few</td>
<td>Own</td>
</tr>
<tr>
<td>Continue a novel</td>
<td>2</td>
<td>Public</td>
<td>Novel</td>
<td>Informal</td>
<td>Mystery</td>
<td>Narrative</td>
<td>Low</td>
<td>Low</td>
<td>T &amp; T</td>
<td>Medium</td>
<td>Few</td>
<td>Other</td>
</tr>
<tr>
<td>Write an article</td>
<td>3</td>
<td>Public</td>
<td>Article</td>
<td>Formal</td>
<td>Society</td>
<td>Descriptive</td>
<td>Low</td>
<td>High</td>
<td>T &amp; T</td>
<td>Medium</td>
<td>Many</td>
<td>Own &amp; others</td>
</tr>
<tr>
<td>Write a speech</td>
<td>4</td>
<td>Public</td>
<td>Speech</td>
<td>Formal</td>
<td>Society</td>
<td>Argumentative</td>
<td>Low</td>
<td>Low</td>
<td>T &amp; T</td>
<td>High</td>
<td>Open</td>
<td>Own &amp; others</td>
</tr>
<tr>
<td>Book review</td>
<td>1-4</td>
<td>Public</td>
<td>Review</td>
<td>Formal</td>
<td>Society</td>
<td>Argumentative</td>
<td>High</td>
<td>High</td>
<td>T &amp; T</td>
<td>High</td>
<td>Open</td>
<td>Own &amp; others</td>
</tr>
</tbody>
</table>

**Table:** Properties and task factors annotated for *Falko Georgetown L2* tasks.
Introduction

Theoretical Background
- Complexity in SLA
- Task Effects in SLA

Data
- Corpora
- Task Annotations

Measuring Complexity
- Automatic System
- Complexity Measures

Descriptive Cross-Corpus Analysis

Inferential Regression Modeling
- Set Up
- Study 1: Task Effects
- Study 2: Performance Effects

Conclusion
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
Overview

- 398 measures of elaborateness and variedness of various domains
- Extracted automatically using elaborate NLP tool chain
- Written by Galasso 2014; Hancke 2013; Weiß 2015, 2017
- Domains:
  1. Language use
  2. Human language processing
  3. Discourse & encoding of meaning
  4. Theoretical linguistics (lexico-semantics, syntax, morphology)
Figure: System pipeline from plain text corpus to feature analysis.
## Resources

<table>
<thead>
<tr>
<th>Task</th>
<th>Component</th>
<th>Version</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenization and sentence segmentation</td>
<td>OpenNLP</td>
<td>1.6.0</td>
<td>default</td>
</tr>
<tr>
<td>POS tagging</td>
<td>Mate tools</td>
<td>3.6.0</td>
<td>default</td>
</tr>
<tr>
<td>Lemmatization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morphological analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependency parsing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compound splitting</td>
<td>JWordSplitter</td>
<td>3.4.0</td>
<td>default</td>
</tr>
<tr>
<td>Constituency parsing</td>
<td>Stanford PCFG parser</td>
<td>3.6.0</td>
<td>default</td>
</tr>
<tr>
<td>Topological field parsing</td>
<td>Berkeley parser</td>
<td>1.7.0</td>
<td>cf. Ramon Ziai</td>
</tr>
</tbody>
</table>

**Table:** NLP components used in the complexity analysis pipeline.
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
Language Use

Domains:
- Corpus and psycho-linguistics

Measures:
- Word frequencies: less frequent = more sophisticated/complex
- Age of acquisition: later AoA = more sophisticated/complex

Implemented:
- AoA approximation based on KCT (Lavalley, Berkling, and Stüker 2015)
Human Language Processing

Domains:
- Cognitive science, psycho-linguistics and information theory

Measures:
- Cognitive processing costs as identified by processing time, reading time, etc.
- Storage and integration of discourse referents consumes cognitive resources
- Long distances between referents increase these costs

Implemented:
- Dependency Locality Theory (DLT) by Gibson 2000; Shain et al. 2016
- Verb argument distances in syllables (Weiß 2015)
Discourse & Encoding of Meaning

Domains:
• Psychology, psycho-linguistics

Measures:
• Propositional idea density (Brown et al. 2008): more propositions = more complex encoding of meaning
• Connectives
• Co-referential expressions
• Grammatical transitions
→ cause more cohesive writing and complex discourse

Implemented:
• PID (Louwerse et al. 2004)
• Connectives as listed by Duden (Gr) 2009
• Local and global overlap of linguistic material
• Co-referential expressions (pronouns, articles, etc.)
• Local transitions of grammatical roles (Barzilay and Lapata 2008; Galasso 2014; Todirascu et al. 2013)
Theoretical Linguistics

Lexio-Semantic
- Measures: concreteness, relatedness, diversity, and variation
- Implemented: TTR, lexical TTR, GermaNet semantic relations

Syntax:
- Measures: clausal complexity (subordination), phrasal complexity (modification)
- Implemented: dependent clause ratios, modifier ratios, complex NP, periphrastic constructions, etc.

Morphology:
- Measures: inflection, derivation, composition
- Implemented: nominalization, tense, compound depth, etc.
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
Overview

- Plots of measures with 95% confidence intervals
- Compare proficiency trajectory across corpora and task profiles
- Sample of over 100 complexity measures
- Grouped under theoretical considerations into concepts
- Selected to represent at least one concept per domain
- All 398 measures at
  http://www.sfs.uni-tuebingen.de/~zweiss/ma-thesis/supplementary-material/complexity-plots/
Human Language Processing
DLT-V and syllable distance measures

Figure: Merlin.

Figure: Falko GT L2; △: curricular tasks, ○: book reviews.
Discourse & Encoding of Meaning

Overlap of linguistic material

Figure: Merlín.

Figure: Falko GT L2; △: curricular tasks, ○: book reviews.
### Syntactic Complexity

#### Complex NPs

<table>
<thead>
<tr>
<th>Component</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words/np</td>
<td>1.6</td>
</tr>
<tr>
<td>NP Deps/npWithDeps</td>
<td>1.5</td>
</tr>
<tr>
<td>NP Mods/npWithDeps</td>
<td>0.6</td>
</tr>
<tr>
<td>Attr Participles/np</td>
<td>0.0</td>
</tr>
<tr>
<td>Clausal NounMods/np</td>
<td>0.0</td>
</tr>
<tr>
<td>Comparative NounMods/np</td>
<td>1.0</td>
</tr>
<tr>
<td>Determiners/np</td>
<td>0.0</td>
</tr>
<tr>
<td>Possessive NounMods/np</td>
<td>0.1</td>
</tr>
<tr>
<td>Prenominal Mods/np</td>
<td>0.2</td>
</tr>
<tr>
<td>Postnominal Mods/np</td>
<td>0.6</td>
</tr>
<tr>
<td>Coverage Modifier Types</td>
<td>2.5</td>
</tr>
</tbody>
</table>

**Figure:** *Merlin.*

**Figure:** *Falko GT L2; △: curricular tasks, ○: book reviews.*
### Morphological Complexity

**Inflection measures**

<table>
<thead>
<tr>
<th>Category</th>
<th>Score 1</th>
<th>Score 2</th>
<th>Score 3</th>
<th>Score 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominatives/noun</td>
<td>0.20</td>
<td>0.25</td>
<td>0.30</td>
<td>0.35</td>
</tr>
<tr>
<td>Accusatives/noun</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>Genitives/noun</td>
<td>0.10</td>
<td>0.15</td>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>Datives/noun</td>
<td>0.70</td>
<td>0.80</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>Vfinal/verb</td>
<td>0.10</td>
<td>0.20</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>InfiniteVerbs/verb</td>
<td>0.04</td>
<td>0.06</td>
<td>0.08</td>
<td>0.10</td>
</tr>
<tr>
<td>ParticipleVerbs/verb</td>
<td>0.00</td>
<td>0.002</td>
<td>0.004</td>
<td>0.006</td>
</tr>
<tr>
<td>Imperatives/vfinal</td>
<td>0.03</td>
<td>0.06</td>
<td>0.09</td>
<td>0.12</td>
</tr>
<tr>
<td>Subjunctives/vfinal</td>
<td>0.85</td>
<td>0.875</td>
<td>0.900</td>
<td>0.925</td>
</tr>
<tr>
<td>Indicatives/vfinal</td>
<td>0.20</td>
<td>0.30</td>
<td>0.40</td>
<td>0.50</td>
</tr>
<tr>
<td>1stPersonInfl/vfinal</td>
<td>0.00</td>
<td>0.025</td>
<td>0.050</td>
<td>0.075</td>
</tr>
<tr>
<td>2ndPersonInfl/vfinal</td>
<td>0.40</td>
<td>0.60</td>
<td>0.80</td>
<td>1.00</td>
</tr>
<tr>
<td>3rdPersonInfl/vfinal</td>
<td>0.25</td>
<td>0.30</td>
<td>0.35</td>
<td>0.40</td>
</tr>
</tbody>
</table>

**Figure:** *Merlin.*

**Figure:** *Falko GT L2; △: curricular tasks, ○: book reviews.*
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
1. Introduction

2. Theoretical Background
   - Complexity in SLA
   - Task Effects in SLA

3. Data
   - Corpora
   - Task Annotations

4. Measuring Complexity
   - Automatic System
   - Complexity Measures

5. Descriptive Cross-Corpus Analysis

6. Inferential Regression Modeling
   - Set Up
     - Study 1: Task Effects
     - Study 2: Performance Effects

7. Conclusion
• Regression analysis to predict proficiency from complexity and task factors
• Use ordinal generative additive regression models (GAMs)
• 2 studies on Merlin: i) task effects; ii) performance effects
• Studies on Falko Georgetown not reported here
Ordinal Generative Additive Regression Models

Overview

GAMs

- Extension of linear regression models
- Use splines as smooths for controlled introduction of non-linear relations
- Highly interpretable, yet similar predictive power as ML techniques like SVM
- Share requirements of regression models: normal, uncorrelated predictors
- Support 1 predictor per 15 to 20 data points

Ordinal Regression

- Link function to non-exponential distribution by Wood 2006
- Estimates boundaries between classes
  → keeps precedence without introducing quantity
Iterative, data-driven model approach

1. Rank measures by information gain using WEKA
2. Test most informative measure for normality; normalize if necessary
3. Test for correlation of predictors
   a. If $< \pm 0.70$ Pearson correlation: add measure to model
   b. Else: remove correlated measures, add measure to model
4. Smooth measures unless they are linear
5. If changes lead to significant model improvement ($\chi^2$ test), keep them
6. Do until 20 iterations did not yield better model or model contains $15/n$ measures
1. Introduction

2. Theoretical Background
   - Complexity in SLA
   - Task Effects in SLA

3. Data
   - Corpora
   - Task Annotations

4. Measuring Complexity
   - Automatic System
   - Complexity Measures

5. Descriptive Cross-Corpus Analysis

6. Inferential Regression Modeling
   - Set Up
     - Study 1: Task Effects
     - Study 2: Performance Effects

7. Conclusion
Study 1: Task Effects

Set Up

Figure: Model formula of *Merlin* interaction model predicting overall CEFR scores from scaled and transformed complexity measures.

```
gam.merlin.interactions <- gam(OverallCefrScore ~
hasTransitionsFromSubjectToNot +
has3rdPersPossessivePronouns +
containsToInfinitives +
usesConjunctinoalClauses +
halfModalClusterPerVP +
logSumNonTerminalNodesPerSentence +
logATFBand2PerTypesFoundInDlex +
avgVTotalIntegrationCostAtFiniteVerb +
lexTypesFoundInDlexPerLexType +
typeTokenRato +
logSumNonTerminalNodesPerWord +
usesConjunctinoalClauses:TaskTheme +
logATFBand2PerTypesFoundInDlex:TaskTheme +
typeTokenRato:TaskTheme +
sumNonTerminalNodesPerWord:TaskTheme +
s(charactersPerWord) +
s(numberOfSentencesSquared) +
TaskTheme,
data=merlin,
family=ocat(R=5))
```
Study 1: Task Effects

Model Fit

- $R^2 = 0.7660$
- Approximately homoscedastic residual errors with $\mu = 0.04; sd = 7.26$ after outlier removal
- Severe outliers across all model variants: Assumption of idiosyncratic properties in texts
- Outliers systematically include learners who performed above or below test level
  - Prompted performance effect analysis in Study 2
### Study 1: Task Effects
#### Model Fit

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>Df</th>
<th>REML</th>
<th>Edf</th>
<th>Compared with</th>
<th>$\chi^2$</th>
<th>Edf difference</th>
<th>$Pr(&gt;\chi^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>1315.05</td>
<td>30.37</td>
<td>658.56</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>1287.08</td>
<td>28.41</td>
<td>642.77</td>
<td>20</td>
<td>Complexity</td>
<td>15.790</td>
<td>1</td>
<td>1.914e-08</td>
</tr>
<tr>
<td>Interaction</td>
<td>1281.00</td>
<td>39.27</td>
<td>628.84</td>
<td>31</td>
<td>Complexity</td>
<td>29.717</td>
<td>12</td>
<td>2.861e-08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reference</td>
<td>13.928</td>
<td>11</td>
<td>0.003</td>
</tr>
</tbody>
</table>

**Table:** Model comparison for complexity, reference, and interaction model build on the *Merlin* data.
# Study 1: Task Effects

## Model Discussion

### A. parametric coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>8.3759</td>
<td>0.3833</td>
<td>21.8509</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td><code>hasTransitionsFromSubjectToNot[TRUE]</code></td>
<td>-0.5349</td>
<td>0.2387</td>
<td>-2.2408</td>
<td>0.0250</td>
</tr>
<tr>
<td><code>has3rdPersPossessivePronouns[TRUE]</code></td>
<td>-0.8906</td>
<td>0.2030</td>
<td>-4.3873</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td><code>containsToInfinitives[TRUE]</code></td>
<td>-0.5541</td>
<td>0.2282</td>
<td>-2.4284</td>
<td>0.0152</td>
</tr>
<tr>
<td><code>halfModalClusterPerVP</code></td>
<td>0.1831</td>
<td>0.1011</td>
<td>1.8113</td>
<td>0.0701</td>
</tr>
<tr>
<td><code>logSumNonTerminalNodesPerSentence</code></td>
<td>1.9714</td>
<td>0.1785</td>
<td>11.0435</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td><code>avgVTotalIntegrationCostAtFiniteVerb</code></td>
<td>0.3705</td>
<td>0.1059</td>
<td>3.4968</td>
<td>0.0005</td>
</tr>
<tr>
<td><code>lexTypesFoundInDlexPerLexType</code></td>
<td>0.8840</td>
<td>0.0942</td>
<td>9.3858</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

**Table**: Interaction model: linear measures.
### Study 1: Task Effects

**Model Discussion**

<table>
<thead>
<tr>
<th>A. parametric coefficients</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>8.3759</td>
<td>0.3833</td>
<td>21.8509</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>usesConjunctionalClauses[TRUE]</td>
<td>-0.6051</td>
<td>0.3173</td>
<td>-1.9074</td>
<td>0.0565</td>
</tr>
<tr>
<td>logATFBand2PerTypesFoundInDlex</td>
<td>-0.3003</td>
<td>0.1091</td>
<td>-2.7528</td>
<td>0.0059</td>
</tr>
<tr>
<td>typeTokenRato</td>
<td>1.2853</td>
<td>0.2038</td>
<td>6.3068</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>logSumNonTerminalNodesPerWord</td>
<td>-0.7130</td>
<td>0.1598</td>
<td>-4.4619</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TaskTheme[Society]</td>
<td>0.4921</td>
<td>0.7085</td>
<td>0.6947</td>
<td>0.4873</td>
</tr>
<tr>
<td>TaskTheme[Profession]</td>
<td>1.0774</td>
<td>0.5508</td>
<td>1.9560</td>
<td>0.0505</td>
</tr>
<tr>
<td>TaskTheme[Smalltalk]</td>
<td>-0.8117</td>
<td>0.3529</td>
<td>-2.3004</td>
<td>0.0214</td>
</tr>
<tr>
<td>usesConjunctionalClauses:TaskTheme[Society]</td>
<td>2.1839</td>
<td>0.9603</td>
<td>2.2742</td>
<td>0.0230</td>
</tr>
<tr>
<td>usesConjunctionalClauses:TaskTheme[Profession]</td>
<td>-0.4185</td>
<td>0.5417</td>
<td>-0.7726</td>
<td>0.4398</td>
</tr>
<tr>
<td>usesConjunctionalClauses:TaskTheme[Smalltalk]</td>
<td>0.5155</td>
<td>0.4714</td>
<td>1.0937</td>
<td>0.2741</td>
</tr>
<tr>
<td>logATFBand2PerTypesFoundInDlex:TaskTheme[Society]</td>
<td>-0.1827</td>
<td>0.4194</td>
<td>-0.4357</td>
<td>0.6631</td>
</tr>
<tr>
<td>logATFBand2PerTypesFoundInDlex:TaskTheme[Profession]</td>
<td>0.5517</td>
<td>0.3530</td>
<td>1.5628</td>
<td>0.1181</td>
</tr>
<tr>
<td>logATFBand2PerTypesFoundInDlex:TaskTheme[Smalltalk]</td>
<td>0.5392</td>
<td>0.2197</td>
<td>2.4539</td>
<td>0.0141</td>
</tr>
<tr>
<td>typeTokenRato:TaskTheme[Society]</td>
<td>-0.4750</td>
<td>0.3634</td>
<td>-1.3072</td>
<td>0.1912</td>
</tr>
<tr>
<td>typeTokenRato:TaskTheme[Profession]</td>
<td>-0.5975</td>
<td>0.3998</td>
<td>-1.4947</td>
<td>0.1350</td>
</tr>
<tr>
<td>typeTokenRato:TaskTheme[Smalltalk]</td>
<td>-0.8335</td>
<td>0.2925</td>
<td>-2.8494</td>
<td>0.0044</td>
</tr>
<tr>
<td>logSumNonTerminalNodesPerWord:TaskTheme[Society]</td>
<td>-0.9369</td>
<td>0.4216</td>
<td>-2.2224</td>
<td>0.0263</td>
</tr>
<tr>
<td>logSumNonTerminalNodesPerWord:TaskTheme[Profession]</td>
<td>-0.1522</td>
<td>0.3409</td>
<td>-0.4465</td>
<td>0.6552</td>
</tr>
<tr>
<td>logSumNonTerminalNodesPerWord:TaskTheme[Smalltalk]</td>
<td>0.2680</td>
<td>0.2344</td>
<td>1.1429</td>
<td>0.2531</td>
</tr>
</tbody>
</table>

**Table:** Interaction model: interactions measures.
Study 1: Task Effects
Model Discussion

B. smooth terms

<table>
<thead>
<tr>
<th>Term</th>
<th>edf</th>
<th>Ref.df</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s(\text{charactersPerWord})$</td>
<td>2.7714</td>
<td>3.5484</td>
<td>18.5670</td>
<td>0.0007</td>
</tr>
<tr>
<td>$s(\text{numberOfSentencesSquared})$</td>
<td>4.6262</td>
<td>5.7193</td>
<td>254.0399</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

**Table:** Interaction model: smoothed measures.

**Figure:** Smooths of *Merlin* interaction model.
## Study 1: Task Effects
### Classification Experiment

<table>
<thead>
<tr>
<th>Model</th>
<th>$\mu$ F1 ±SD</th>
<th>$\mu$ Recall ±SD</th>
<th>$\mu$ Precision ±SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>7.37 11.59</td>
<td>7.44 11.33</td>
<td>7.37 11.37</td>
</tr>
<tr>
<td>Complexity</td>
<td>70.97 4.25</td>
<td>71.63 4.74</td>
<td>72.30 4.09</td>
</tr>
<tr>
<td>Reference</td>
<td>71.32 4.33</td>
<td>71.78 4.87</td>
<td>72.74 4.10</td>
</tr>
<tr>
<td>Interaction</td>
<td>72.17 4.43</td>
<td>72.69 4.94</td>
<td>73.39 4.15</td>
</tr>
</tbody>
</table>

**Table**: Weighted average precision, recall, and f1 score for complexity, reference, and interaction model for 10 iterations of 10-folds cross-validation.
### Study 1: Task Effects

#### Classification Experiment

<table>
<thead>
<tr>
<th>Predicted↓ / Observed→</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>25.5</td>
<td>10.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A2</td>
<td><strong>29.5</strong></td>
<td><strong>241.3</strong></td>
<td>45.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>B1</td>
<td>0.0</td>
<td>52.6</td>
<td><strong>233.5</strong></td>
<td>37.8</td>
<td>0.0</td>
</tr>
<tr>
<td>B2</td>
<td>0.0</td>
<td>0.0</td>
<td>49.1</td>
<td><strong>243.1</strong></td>
<td><strong>37.9</strong></td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>10.1</td>
<td>8.1</td>
</tr>
</tbody>
</table>

**Table:** Averaged confusion matrix for classification of L2 proficiency in *Merlin* using the interaction model.
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
Study 2: Performance Effects

Set Up

**Figure:** Model formula of *Merlin* success model predicting overall CEFR scores from scaled and transformed complexity measures

```r
506  gam.merlin.success <- gam(OverallCefrScore ~
507    hasTransitionsFromSubjectToNot +
508    has3rdPersPossessivePronouns +
509    containsToInfinitives +
510    usesConjunctinalClauses +
511    logSumNonTerminalNodesPerSentence +
512    logATFBand2PerTypesFoundInDlex +
513    avgVTotalIntegrationCostAtFiniteVerb +
514    lexTypesFoundInDlexPerLexType +
515    typeTokenRato +
516    logSumNonTerminalNodesPerWord +
517    logATFBand2PerTypesFoundInDlex:TaskTheme +
518    s(charactersPerWord, by = Passed, k = 6) +
519    s(numberOfSentencesSquared, by = Passed) +
520    Passed +
521    TaskTheme,
522    data=merlin,
523    family=ocat(R=n_cat))
```
Study 2: Performance Effects

Model Fit

- $R^2 = 0.9000$
- Approximately homoscedastic residual errors with $\mu = -0.14; sd = 12.91$ after outlier removal
- Same outliers as before, except for under-performing learners
- Outliers still systematically include learners who performed above or below test level
Study 2: Performance Effects

Model Fit

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>Df</th>
<th>REML</th>
<th>Edf</th>
<th>Comparison with</th>
<th>$\chi^2$</th>
<th>Edf diff</th>
<th>$Pr(&gt; \chi^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>1315.05</td>
<td>30.37</td>
<td>658.56</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference model</td>
<td>1287.08</td>
<td>28.41</td>
<td>642.77</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction model</td>
<td>1281.00</td>
<td>39.27</td>
<td>628.84</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success model</td>
<td>821.11</td>
<td>35.76</td>
<td>401.19</td>
<td>26</td>
<td>Complexity</td>
<td>257.36</td>
<td>7</td>
<td>$&lt; 2e - 16$</td>
</tr>
<tr>
<td>Success model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reference</td>
<td>241.573</td>
<td>6</td>
<td>$&lt; 2e - 16$</td>
</tr>
<tr>
<td>Success model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Interaction</td>
<td>227.65</td>
<td>-5</td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Model comparison for reference, complexity, interaction, and success GAMs modeling L2 proficiency from complexity measures and task theme on the *Merlin* data.
### Study 2: Performance Effects

#### Model Discussion

#### A. parametric coefficients

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.8610</td>
<td>0.5186</td>
<td>7.4455</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>hasTransitionsFromSubjectToNot[TRUE]</td>
<td>-0.8565</td>
<td>0.2732</td>
<td>-3.1350</td>
<td>0.0017</td>
</tr>
<tr>
<td>has3rdPersPossessivePronouns[TRUE]</td>
<td>-1.3267</td>
<td>0.2556</td>
<td>-5.1903</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>containsToInfinitives[TRUE]</td>
<td>-0.7246</td>
<td>0.2701</td>
<td>-2.6825</td>
<td>0.0073</td>
</tr>
<tr>
<td>usesConjuncturalClauses[TRUE]</td>
<td>-0.5514</td>
<td>0.2698</td>
<td>-2.0434</td>
<td>0.0410</td>
</tr>
<tr>
<td>logATFBand2PerTypesFoundInDlex</td>
<td>-0.3972</td>
<td>0.1202</td>
<td>-3.3054</td>
<td>0.0009</td>
</tr>
<tr>
<td>avgVTotalIntegrationCostAtFiniteVerb</td>
<td>0.4765</td>
<td>0.1380</td>
<td>3.4522</td>
<td>0.0006</td>
</tr>
<tr>
<td>lexTypesFoundInDlexPerLexType</td>
<td>0.9649</td>
<td>0.1132</td>
<td>8.5218</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>typeTokenRato</td>
<td>1.2797</td>
<td>0.1877</td>
<td>6.8176</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>sumNonTerminalNodesPerWord</td>
<td>-0.8316</td>
<td>0.1398</td>
<td>-5.9464</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>logSumNonTerminalNodesPerSentence</td>
<td>2.4829</td>
<td>0.2093</td>
<td>11.8655</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Passed[TRUE]</td>
<td>6.6843</td>
<td>0.3018</td>
<td>22.1510</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TaskTheme[Society]</td>
<td>11.5649</td>
<td>0.6668</td>
<td>17.3437</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TaskTheme[Profession]</td>
<td>7.2479</td>
<td>0.5982</td>
<td>12.1158</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TaskTheme[Smalltalk]</td>
<td>0.9101</td>
<td>0.2796</td>
<td>3.2550</td>
<td>0.0011</td>
</tr>
<tr>
<td>logATFBand2PerTypesFoundInDlex:TaskTheme[Society]</td>
<td>0.4881</td>
<td>0.5980</td>
<td>0.8163</td>
<td>0.4144</td>
</tr>
<tr>
<td>logATFBand2PerTypesFoundInDlex:TaskTheme[Profession]</td>
<td>1.1930</td>
<td>0.4812</td>
<td>2.4795</td>
<td>0.0132</td>
</tr>
<tr>
<td>logATFBand2PerTypesFoundInDlex:TaskTheme[Smalltalk]</td>
<td>0.3248</td>
<td>0.2428</td>
<td>1.3376</td>
<td>0.1810</td>
</tr>
<tr>
<td>s(charactersPerWord):Passed[FALSE]</td>
<td>2.5964</td>
<td>3.2239</td>
<td>5.3214</td>
<td>0.1503</td>
</tr>
<tr>
<td>s(charactersPerWord):Passed[TRUE]</td>
<td>1.3498</td>
<td>1.6284</td>
<td>6.5613</td>
<td>0.0297</td>
</tr>
<tr>
<td>s(numberOfSentencesSquared):Passed[FALSE]</td>
<td>3.6322</td>
<td>4.5433</td>
<td>69.6021</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>s(numberOfSentencesSquared):Passed[TRUE]</td>
<td>4.3517</td>
<td>5.3657</td>
<td>306.2779</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

**Table:** Summary of success model predicting Merlin overall CEFR scores from scaled and transformed complexity measures in Merlin. Uses ‘demand’ as reference level.
Study 2: Performance Effects
Model Discussion

Figure: Smooths of Merlin success model.
Study 2: Performance Effects
Model Discussion

- Most task theme interactions become uninformative
- Still significantly different slopes, but not enough new variance explained
- Especially: texts about society heavily confounded with failed tests
- Unclear relationship between performance and task theme
## Study 2: Performance Effects

### Classification Experiment

<table>
<thead>
<tr>
<th>Model</th>
<th>$\mu$ F1</th>
<th>$\pm$ SD</th>
<th>$\mu$ Recall</th>
<th>$\pm$ SD</th>
<th>$\mu$ Precision</th>
<th>$\pm$ SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>7.37</td>
<td>11.59</td>
<td>7.44</td>
<td>11.33</td>
<td>7.37</td>
<td>11.37</td>
</tr>
<tr>
<td>Complexity</td>
<td>71.20</td>
<td>4.25</td>
<td>71.89</td>
<td>4.71</td>
<td>72.53</td>
<td>4.03</td>
</tr>
<tr>
<td>Reference</td>
<td>71.32</td>
<td>4.33</td>
<td>71.78</td>
<td>4.87</td>
<td>72.74</td>
<td>4.10</td>
</tr>
<tr>
<td>Interaction</td>
<td>72.17</td>
<td>4.43</td>
<td>72.69</td>
<td>4.94</td>
<td>73.39</td>
<td>4.15</td>
</tr>
<tr>
<td>Success</td>
<td>84.98</td>
<td>2.75</td>
<td>85.60</td>
<td>2.80</td>
<td>85.28</td>
<td>2.74</td>
</tr>
</tbody>
</table>

**Table:** Weighted average precision, recall, and F1 score for complexity, reference, interaction and success model for 10 iterations of 10-folds cross-validation.
### Study 2: Performance Effects
Classification Experiment

<table>
<thead>
<tr>
<th>Pred.↓ / Obs.→</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td><strong>28.1</strong></td>
<td>12.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A2</td>
<td>26.9</td>
<td><strong>260.9</strong></td>
<td>34.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>B1</td>
<td>0.0</td>
<td>30.4</td>
<td><strong>271.9</strong></td>
<td>18.2</td>
<td>0.0</td>
</tr>
<tr>
<td>B2</td>
<td>0.0</td>
<td>0.0</td>
<td>21.6</td>
<td><strong>271.9</strong></td>
<td>6.0</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.4</td>
<td><strong>40.0</strong></td>
</tr>
</tbody>
</table>

**Table:** Averaged confusion matrix for classification of L2 proficiency in *Merlin* using the success model.
1 Introduction

2 Theoretical Background
   Complexity in SLA
   Task Effects in SLA

3 Data
   Corpora
   Task Annotations

4 Measuring Complexity
   Automatic System
   Complexity Measures

5 Descriptive Cross-Corpus Analysis

6 Inferential Regression Modeling
   Set Up
   Study 1: Task Effects
   Study 2: Performance Effects

7 Conclusion
Conclusion

Findings

How do measures of complexity model German L2 proficiency?

- Most indices of the same concept tend to develop homogeneously and stable across corpora
- Most indices develop homogeneously across corpora
- Data-driven feature selection approaches yield diverse set of measures
- GAMs are highly interpretable, yet show considerable predictive power
To which extend is this influenced by cognitive or functional task-effects?

- Some measures are more stable across heterogeneous task backgrounds (human language processing, complex NPs)
- Other measures are less stable
- Stable measures especially promising for systems evaluating diverse task backgrounds
- Task factors seem to predominantly effect local measures of structural complexity
- Further research on this required
Conclusion

Findings

Does a retrospective analysis of German learner corpora with diverse task backgrounds improve complexity-based L2 proficiency modeling?

- Post-hoc annotation straightforward if task documentation available
- Suited to decrease confound of tasks and course levels
- Task factors improve model fit significantly and decrease non-linearity
- Interactions seem unstable, models suffer from wide standard deviation
- Results lack interpretability due to skewed distribution
- Analysis improves situation, but idiosyncratic distributional properties of data remain problematic
Future Work

Next:

1. Investigation of task and performance effects on more balanced data sets
2. Study of adequacy of L2 complexity by comparing results on comparable L1 productions (Falko)
3. Make complexity code used here publicly available in CTAP (Chen and Meurers 2016)

Also interesting:

- Analysis of task type interactions in data
- Cross-corpus testing of success model
- Systematically assess sensitivity of system and structure complexity to task effects
- Systematic validation of measure validity on L2 data
Thank you for your attention! Questions?


References II


References III


Merlin project (2014f). *task description: Informal e-mail: arrange an appointment with a friend to go swimming together*. http://merlin-platform.eu/.

Merlin project (2014g). *task description: Informal e-mail: ask a friend for help with finding an apartment*. http://merlin-platform.eu/.


References VII


## Robinson’s Cognition Hypothesis

<table>
<thead>
<tr>
<th>Task complexity</th>
<th>Task conditions</th>
<th>Task difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>(cognitive factors)</td>
<td>(interational factors)</td>
<td>(learner factors)</td>
</tr>
<tr>
<td>(a) resource-directing</td>
<td>(a) participation variables</td>
<td>(a) affective variables</td>
</tr>
<tr>
<td>e.g., +/- few elements</td>
<td>e.g., open/closed</td>
<td>e.g., motivation</td>
</tr>
<tr>
<td>+/- here-and-now</td>
<td>one-way/two-way</td>
<td>anxiety</td>
</tr>
<tr>
<td>+/- no reasoning demands</td>
<td>convergent/divergent</td>
<td>confidence</td>
</tr>
<tr>
<td>(b) resource-depleting</td>
<td>(b) participant variables</td>
<td>(b) ability variables</td>
</tr>
<tr>
<td>e.g., +/- planning</td>
<td>e.g., gender</td>
<td>e.g., aptitude</td>
</tr>
<tr>
<td>+/- single task</td>
<td>familiarity</td>
<td>proficiency</td>
</tr>
<tr>
<td>+/- prior knowledge</td>
<td>power/solidarity</td>
<td>intelligence</td>
</tr>
</tbody>
</table>

Sequencing criteria
- Prospective decisions about task units

Methodological influences
- On-line decisions about pairs and groups (based on Robinson, in pressa)

**Figure:** Task complexity, condition, and difficulty (Robinson 2001, p. 30, Figure 1).
Language Use
DlexDB frequencies

Figure: Merlin.

Figure: Falko GT L2; △: curricular tasks, ○: book reviews.
Discourse & Encoding of Meaning

Pronouns, articles and names

Figure: Merlin.

Figure: Falko GT L2; △: curricular tasks, ○: book reviews.
Lexical Complexity

Lexical Variation

Figure: *Merlin.*

Figure: *Falko GT L2; △: curricular tasks, ○: book reviews.*
## Syntactic Complexity

### Periphrastic grammatical measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Eventive Passive/Fin Clause</th>
<th>Passive/Fin Clause</th>
<th>Quasi Passive/Fin Clause</th>
<th>Sein/Verbs</th>
<th>Haben/Verbs</th>
<th>Simple Present/VFin</th>
<th>Simple Past/VFin</th>
<th>Present Perfect/VFin</th>
<th>Past Perfect/VFin</th>
<th>Future 1/VFin</th>
<th>Future 2/VFin</th>
<th>Coverage Tenses</th>
<th>Coverage Periphrastic Tenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure:** *Merlin.*

**Figure:** *Falko GT L2; △: curricular tasks, ○: book reviews.*
Syntactic Complexity

Dependent clause measures

Figure: Merlin.

Figure: Falko GT L2; △: curricular tasks, ○: book reviews.
Generative Additive Regression Models
From Linear to Additive Models

\[ \hat{y} = \eta + \epsilon, \]  
where \( \epsilon \sim N(0, \sigma^2) \) and \( \eta = \beta_0 + \sum_{i=1}^{l} x_i \beta_i \) \hspace{1cm} (1)

\[ g(\hat{y}) = \eta + \epsilon, \]  
where \( \eta = \beta_0 + \sum_{i=1}^{l} x_i \beta_i \) \hspace{1cm} (2)

\[ g(\hat{y}) = \eta + \epsilon, \]  
where \( \eta = \beta_0 + \sum_{i=1}^{l} s_i(x_i) \) \hspace{1cm} (3)
Generative Additive Regression Models
From Linear to Additive Models

\[ g(\hat{y}) = \eta + \epsilon, \text{ where } \eta = \beta_0 + \sum_{i=1}^{l} s_i(x_i) \]  \hspace{1cm} (4)

\[ s(x) = \sum_{k=1}^{K} b_k(x) \beta_k, \] \hspace{1cm} (5)

\[ s(x) = \sum_{c=1}^{C+1} x^{c-1} \beta_c \] \hspace{1cm} (6)
Generative Additive Regression Models

Regression Splines

Figure: Single cubic basis function (left) and full cubic regression spline (right), cf. Wood 2006, p. 147, Figure 4.1.
Generative Additive Regression Models
Regression Splines

Figure: A rank 7 thin plate regression spline preceded by its weighted basis functions, cf. Wood 2006, p. 153, Figure 4.5.
Generative Additive Regression Models

Ordinal Models

\[ u = \eta + \epsilon, \text{ where } \eta = \beta_0 + \sum_{i=1}^{l} s_i(x_i), \text{ and } u \in [\pm \infty] \]  

\( (7) \)

- Ordinal data neither numeric nor nominal
- Ordinal distribution not covered in exponential link functions \( g() \)
- Solution by Wood 2006: partition \( \pm \infty \) into \( K \) bins using \( K-1 \) boundaries
- Estimate latent variable \( u \) with regression model
- Assign ordinal category based on interval in which \( u \) falls
Generative Additive Regression Models

Ordinal Models

Figure: Mapping of latent variable $u$ to CEFR levels using estimated boundaries from the *interaction* model of Study 1.
Automatic System
Overview

English Systems
- Coh-Metrix Web Interface (McNamara et al. 2014)
- Syntactic and Lexical Complexity analyzer (Lu 2010)
- Linguistic Analysis Tool (Kyle 2016)
- Common Text Analysis Platform (Chen and Meurers 2016)

German Systems
- DeLite Readability Checker (von der Brück and Hartrumpf 2007; von der Brück, Hartrumpf, and Helbig 2008)
- Tübinger complexity code (Hancke, Vajjala, and Meurers 2012; Weiß 2015)