Instructed second language acquisition and longitudinal learner corpus research: The case of lexical and syntactic complexity

Nina Vyatkina
University of Kansas

Hagen Hirschmann,
Felix Golcher
Humboldt-Universität zu Berlin

TaLC XII
Giessen, July 21, 2016
Overview

• Research goal:
  – Map development of L2 lexical complexity onto development of syntactic complexity explored in our earlier study

• Methodological question:
  – How can we describe the development of L2 writing complexity in early learners in an instructed setting?
Theoretical background

• Usage-Based Grammar
  – languages are learned primarily bottom-up: from specific examples to low-scope patterns to abstract constructions
  – inseparability of grammar and the lexicon
    Bybee 2008; Ellis 2014; Flowerdew, 2011; Langacker 1987; Ortega 2015; Robinson & Ellis 2008

• Dynamic Systems Theory
  – L2 development is a dynamic process, in which regular growth stages are modulated by a complex variation within and among individuals as well as interrelated aspects of the interlanguage system
    Larsen-Freeman 2006; Verspoor et al. 2008
L2 Complexity

• Measuring learner progress and proficiency – indicators employed in SLA since 1980s (Larsen-Freeman, 1983; Skehan, 1989)

• → CAF Measures:
  – Complexity:
    • the extent to which the language produced in performing a task is elaborate and varied (Ellis, 2003)
    • the range of forms that surface in language production and the degree of sophistication of such forms (Ortega, 2003)
  – Accuracy: error-free L2 production
  – Fluency: speed of L2 production
L2 writing complexity research

• Primarily explored **structural** measures of syntactic and lexical complexity:
  – Syntactic complexity: length and ratios of syntactic units
    • words, clauses / sentences, T-units...
  – Lexical complexity: ratios measuring word diversity, density, and sophistication
    • type-token ratios, content words/functional words, rare words/common words, ...

• Research syntheses: linear increase in some but not all measures with increasing proficiency; complex interactions between measures
Designs

• Many complexity studies:
  – cross-sectional or single-case longitudinal
  – manual annotation of selected features

• This study:
  – longitudinal corpus, multiple learner profiles
  – automatic corpus-based profiling (POS and lemma annotation)

Granger & Rayson 1998; Hawkins & McCarthy 2010; Ortega & Sinicrope 2008
Data: subset of KANDEL

KANDEL is a pos-annotated, lemmatized, and error-annotated open access learner corpus


This study: longitudinal KANDEL subset

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>setting</td>
<td>Instructed SLA, large public US university</td>
</tr>
<tr>
<td>participants</td>
<td>12 students (5 male, 7 female)</td>
</tr>
<tr>
<td>age</td>
<td>18-22 (mean 19.5), 1 learner &gt;30</td>
</tr>
<tr>
<td>languages</td>
<td>L1 English, L2 German (beginner to A2 CEFR proficiency)</td>
</tr>
<tr>
<td>time</td>
<td>4 semesters, 17 data collection points (every 3-5 weeks)</td>
</tr>
<tr>
<td>texts</td>
<td>185 rough drafts in-class and at-home L2 essays (personal narratives and descriptions; essays with explanatory elements; letters)</td>
</tr>
<tr>
<td>text length</td>
<td>100-200 words (mean 161)</td>
</tr>
</tbody>
</table>
Research question and hypothesis

• RQ: Does the observed development of specific word classes (syntactic modifiers) correlate with lexical development?
• RH: Lexical richness is verifiably increasing over time, independently of growth curve of syntactic categories
Lexical complexity measures

• Structural measures:
  – Lexical density
  – Lexical sophistication
  – Lexical diversity (TTR and type frequency)

• Content-based measures:
  – lexical novelty (emergent words)
  – specific content words as specific syntactic modifiers (cf. Ortega & Sinicrope 2008)
KanDeL in ANNIS – sample search
KanDeL in ANNIS – sample search
Procedure

• Focusing on modifier categories
  1. 'prenominal adjective',
  2. 'predicative adjective',
  3. 'adverb'
     – very general categories, contain different syntactic and semantic types

• Processing steps for study:
  – Export all relevant tokens with sentence contexts
  – Annotating individual tokens in MS Excel: functional syntactic and semantic categories
     • Excluding all erroneous tokens that cannot be interpreted (orthographic vs. grammatical errors)
  – Data analyses using R and MS Excel tables
Procedure

- Focusing on modifier categories
  1. 'prenominal adjective',
  2. 'predicative adjective',
  3. 'adverb'
  - very general categories, contain different syntactic and semantic types

<table>
<thead>
<tr>
<th>1</th>
<th>tok</th>
<th>lemma</th>
<th>lemma_ZH</th>
<th>pos_corr</th>
<th>name</th>
<th>topic</th>
<th>zeitpunkt</th>
<th>Freq lemma</th>
<th>ADV func</th>
<th>Semantik</th>
</tr>
</thead>
<tbody>
<tr>
<td>440</td>
<td>noch</td>
<td>noch</td>
<td>noch</td>
<td>ADV</td>
<td>Ellen</td>
<td>Tipps für ein</td>
<td>11</td>
<td>13</td>
<td>Adv</td>
<td>Temp</td>
</tr>
<tr>
<td>441</td>
<td>noch</td>
<td>noch</td>
<td>noch</td>
<td>ADV</td>
<td>Ellen</td>
<td>Tipps für ein</td>
<td>11</td>
<td>13</td>
<td>Adv</td>
<td>Temp</td>
</tr>
<tr>
<td>442</td>
<td>noch</td>
<td>noch</td>
<td>noch</td>
<td>ADV</td>
<td>Jessica</td>
<td>Tipps für ein</td>
<td>11</td>
<td>13</td>
<td>Ptkint</td>
<td>Int</td>
</tr>
<tr>
<td>443</td>
<td>noch</td>
<td>noch</td>
<td>noch</td>
<td>ADV</td>
<td>Ramona</td>
<td>Tipps für ein</td>
<td>11</td>
<td>13</td>
<td>Adv</td>
<td>Temp</td>
</tr>
<tr>
<td>444</td>
<td>noch</td>
<td>noch</td>
<td>noch</td>
<td>ADV</td>
<td>Robert</td>
<td>Tipps für ein</td>
<td>11</td>
<td>13</td>
<td>Adv</td>
<td>Temp</td>
</tr>
<tr>
<td>445</td>
<td>nur</td>
<td>nur</td>
<td>nur</td>
<td>ADV</td>
<td>Elyse</td>
<td>Tipps für ein</td>
<td>11</td>
<td>20</td>
<td>PtkFo</td>
<td>nur</td>
</tr>
<tr>
<td>446</td>
<td>Schliesslich</td>
<td>schliesslich</td>
<td>schliesslich</td>
<td>ADV</td>
<td>Ellen</td>
<td>Tipps für ein</td>
<td>11</td>
<td>1</td>
<td>Adv</td>
<td>Temp</td>
</tr>
<tr>
<td>447</td>
<td>schon</td>
<td>schon</td>
<td>schon</td>
<td>ADV</td>
<td>Sophia</td>
<td>Tipps für ein</td>
<td>11</td>
<td>6</td>
<td>ERR</td>
<td></td>
</tr>
<tr>
<td>448</td>
<td>sehr</td>
<td>sehr</td>
<td>sehr</td>
<td>ADV</td>
<td>Aimon</td>
<td>Tipps für ein</td>
<td>11</td>
<td>149</td>
<td>Ptkint</td>
<td>Int</td>
</tr>
<tr>
<td>449</td>
<td>sehr</td>
<td>sehr</td>
<td>sehr</td>
<td>ADV</td>
<td>Aimon</td>
<td>Tipps für ein</td>
<td>11</td>
<td>149</td>
<td>Ptkint</td>
<td>Int</td>
</tr>
<tr>
<td>450</td>
<td>sehr</td>
<td>sehr</td>
<td>sehr</td>
<td>ADV</td>
<td>Elyse</td>
<td>Tipps für ein</td>
<td>11</td>
<td>149</td>
<td>Ptkint</td>
<td>Int</td>
</tr>
<tr>
<td>451</td>
<td>sehr</td>
<td>sehr</td>
<td>sehr</td>
<td>ADV</td>
<td>Elyse</td>
<td>Tipps für ein</td>
<td>11</td>
<td>149</td>
<td>Ptkint</td>
<td>Int</td>
</tr>
<tr>
<td>452</td>
<td>sehr</td>
<td>sehr</td>
<td>sehr</td>
<td>ADV</td>
<td>Ivan</td>
<td>Tipps für ein</td>
<td>11</td>
<td>149</td>
<td>Ptkint</td>
<td>Int</td>
</tr>
<tr>
<td>453</td>
<td>sehr</td>
<td>sehr</td>
<td>sehr</td>
<td>ADV</td>
<td>Jade</td>
<td>Tipps für ein</td>
<td>11</td>
<td>149</td>
<td>Ptkint</td>
<td>Int</td>
</tr>
<tr>
<td>454</td>
<td>sehr</td>
<td>sehr</td>
<td>sehr</td>
<td>ADV</td>
<td>Jessica</td>
<td>Tipps für ein</td>
<td>11</td>
<td>149</td>
<td>Ptkint</td>
<td>Int</td>
</tr>
</tbody>
</table>
Use of modifier categories...
(Vyatkina & Hirschmann & Golcher 2015, no lexical perspective)

• **Prenominal adjectives** significantly increasing over time (despite great variation)

• **Predicative adjectives** significantly decreasing over time (despite great variation)

• **Adverbs** show no significant trend

*Ich habe die **beste** Familie in der Welt.* (Aimon 03)
*I have the **best** family in the world.*

*Sie ist sehr **schön**.* (Aimon 03)
*She is very **pretty**.*

*Gestern kam Julchen zu mir.* (Patrick 15)
*Yesterday came Julchen to me.*
Use of modifier categories...
(Vyatkina&Hirschmann&Golcher 2015, no lexical perspective)

- **Prenominal adjectives** significantly increasing over time (despite great variation)
- **Predicative adjectives** significantly decreasing over time (despite great variation)
- **Adverbs** show no significant trend

Ich habe die *beste* Familie in der Welt. (Aimon 03)
I have the *best* family in the world.

Sie ist *sehr schön*. (Aimon 03)
She is *very pretty*.

Gestern kam Julchen zu mir. (Patrick 15)
Yesterday came Julchen to me.
Results: TTR

- TTR over time: black dots $\rightarrow$ TTR per text and point in time (bigger dots symbolize longer texts)
Results: TTR

- TTR over time: black dots $\rightarrow$ TTR per text and point in time (bigger dots symbolize longer texts)

- short texts $\rightarrow$ many single occurrences per point in time
- few frequent occurrences $\rightarrow$ no clear development
Results: Types per text

- Types over time: black dots → absolute type frequency per text and point in time, **black** dots: individual texts, **blue** dots: mean type values for group, **green** dots: mean token values
Results: Types per text

- Types over time: black dots \(\rightarrow\) absolute type frequency per text and point in time, black dots: individual texts, blue dots: mean type values for group, green dots: mean token values
Results: New types

- New types per point in time and individual person. Red and blue dots: single texts, black dots: mean values for group with bootstrapped confidence intervals
Results: New types

- New types per point in time and individual person. Red and blue dots: single texts, black dots: mean values for group with bootstrapped confidence intervals.

⇒ Relative frequencies of new occurrences per point in time goes hand in hand with categorial use.
Individual lexemes per point in time (lexical diversity for whole group)

- **ADJA**

- **ADJD**

- **ADV**

  most frequent:

<table>
<thead>
<tr>
<th>Lexeme</th>
<th>Frequency</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>sehr (very)</td>
<td>15,6</td>
<td>149</td>
</tr>
<tr>
<td>auch (also)</td>
<td>9,6</td>
<td>92</td>
</tr>
<tr>
<td>gern (with pleasure)</td>
<td>7,0</td>
<td>67</td>
</tr>
<tr>
<td>jetzt (now)</td>
<td>4,2</td>
<td>40</td>
</tr>
<tr>
<td>aber (however)</td>
<td>4,0</td>
<td>38</td>
</tr>
</tbody>
</table>

  most frequent:

<table>
<thead>
<tr>
<th>Lexeme</th>
<th>Frequency</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>gut (good)</td>
<td>10,5</td>
<td>65</td>
</tr>
<tr>
<td>neu (new)</td>
<td>6,5</td>
<td>40</td>
</tr>
<tr>
<td>jung (young)</td>
<td>6,5</td>
<td>40</td>
</tr>
<tr>
<td>erst (first)</td>
<td>4,1</td>
<td>25</td>
</tr>
<tr>
<td>silber (silver)</td>
<td>3,0</td>
<td>18</td>
</tr>
</tbody>
</table>

  most frequent:

<table>
<thead>
<tr>
<th>Lexeme</th>
<th>Frequency</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>gut (good)</td>
<td>9,7</td>
<td>83</td>
</tr>
<tr>
<td>groß (big)</td>
<td>4,0</td>
<td>34</td>
</tr>
<tr>
<td>deutsch (German)</td>
<td>3,4</td>
<td>29</td>
</tr>
<tr>
<td>interessant (interesting)</td>
<td>3,0</td>
<td>26</td>
</tr>
</tbody>
</table>
Individual lexemes per point in time (lexical diversity for whole group)

- ADJA

- ADJD

- ADV

most frequent:

- gut (good) 10,5 (65)
- neu (new) 6,5 (40)
- jung (young) 6,5 (40)
- erst (first) 4,1 (25)
- silbern (silver) 3,0 (18)

most frequent:

- gut (good) 9,7 (83)
- groß (big) 4,0 (34)
- deutsch (German) 3,4 (29)
- interessant (interesting) 3,0 (26)

most frequent:

- sehr (very) 15,6 (149)
- auch (also) 9,6 (92)
- gern (with pleasure) 7,0 (67)
- jetzt (now) 4,2 (40)
- aber (however) 4,0 (38)

\[no\ clear\ lexical\ diversification\ over\ time\]
Results for ADV

Use of category adverb (ADV) according to Vyatkina & Hirschmann & Golcher 2015

- Now taking a look at heterogeneous category ADV:
  - Variation of lexeme use within group
  - Frequencies of adverb subcategories
ADV lexemes used by number x of learner per point in time
ADV lexemes used by number x of learner per point in time

- Few lexemes used by many students
- Few lexemes used at many points of time
Results: ADV – semantic categories

- Semantic categories: temporal, intensifying, locative, evaluative, modal, epistemic, focus modification, causal, adversative
Results: ADV – semantic categories

Lexical winners:
- dann-then
- sehr-very
- gern-with pleasure
- hier-here

Semantic categories: temporal, intensifying, locative, evaluative, modal, epistemic, focus modification, causal, adversative
Results: ADV – semantic categories

Lexical winners:
- dann-then
- sehr-very
- gern-with pleasure
- hier-here

→ Slight diversification of semantic types over time visible
   (the same is true for functional categories
   like "sentence adverbial")

- Semantic categories: temporal, intensifying, locative, evaluative, modal, epistemic, focus modification, causal, adversative
Conclusions: correlations between lexical and syntactic measures

- Longitudinal KANDEL data allows for qualitative and quantitative descriptions of learner development
- Correlation of lexical "concepts" with categorial use in KANDEL data: new types per point in time > types per point of time > TTR per point of time
- Generally, less systematic growth of lexical diversity than expected → Hypothesis "RH" not confirmed
- Huge individual differences (despite homogeneous learner group)
- But systematic developments on different grammatical levels:
  - semantic categories: adversative and causal adverbs
  - functional categories: sentence adverbs and modal particles
- ‘lexical teddybears’ in many subclasses (e.g. sehr—‘very’ – an absolute winner for intensifiers)
- Task and topic effects observed especially on semantic level
Future research directions

• Correlations between complexity and accuracy
• Analysis of lexico-grammatical constructions
• Analysis of pseudo-longitudinal (cohort) data – much larger KANDEL subcorpora
Thanks for your comments!

Acknowledgments:
• Marc Reznicek (for contributing to earlier stages of this study)
• Emily Hackmann and Michael Grünbaum (University of Kansas) for data annotation
• The German and American Fulbright Commission
• KU Institute for Digital Research in the Humanities, Digital Humanities Seed Grant
• Language Learning Small Grants Research Program

Nina Vyatkina  vyatkina@ku.edu
Hagen Hirschmann  hirschhx@hu-berlin.de
Felix Golcher  felix.golcher@hu-berlin.de