

# Continuation writing shows alignment in linguistic complexity

Xiaobin Chen and Detmar Meurers

ICALL Research Group

LEAD Graduate School & Research Network

University of Tübingen

“At the Interface of Second Language Acquisition and Computational Linguistics”  
INDUS Meeting Tübingen 28.2.-1.3.2018

- ▶ Comprehensible input fosters acquisition
  - ▶  $i+1$  (Krashen 1985)
  - ▶ Zone of Proximal Development (Vygotsky 1986)
- ▶ Questions:
  - ▶ How can we characterize the interlanguage “ $i$ ”?
    - ▶ How can we observe and quantify different aspects of the interlanguage?
  - ▶ How can we determine the challenge “ $+1$ ”?
    - ▶ How much challenge in terms of which aspects of the language, for which individual learner?
- ▶ Our approach: Analyze linguistic complexity
  - ▶ of **learner productions** as characteristic of the individual learner’s interlanguage
  - ▶ of **input** for the learner as characteristic of the challenge

## Introduction

Relating input to  
production using  
complexity analysis

Results

Findings

Application

Outlook

References

- ▶ Definition: variedness and elaborateness of language  
(Rescher 1998; Ellis 2003)
- ▶ Complexity observable at all levels of the linguistic system  
(lexis, morphology, syntax, semantics, discourse) and  
human sentence processing
- ▶ Linguistic complexity is used to characterize language  
development of learners:
  - ▶ Complexity, Accuracy, Fluency (CAF, Housen et al. 2012)
- ▶ Our idea: empirically explore the  $i+1$  in terms of the  
complexity of learner productions and learner input

Continuation Writing corpus (Wang & Wang 2015).

- ▶ Participants: 48 adult Chinese EFL learners
- ▶ Task: continuation writing of stories with endings removed
- ▶ Conditions: continuation writing of stories after reading
  - (i) a Chinese text: **baseline writing**
  - (ii) an English text: **continuation writing**
- ▶ Wang & Wang (2015) results:
  - ▶ significantly fewer linguistic errors in Condition (ii)
  - ▶ interpreted as **alignment with input**

- ▶ Feature extraction: 558 complexity measures for both the input texts and student writings
  - ▶ Lexical: counts, lexical variation indexes, lexical sophistication measures...
  - ▶ Syntactic: component counts, complex structure indexes, edit distances of syntactic trees...
  - ▶ Cohesion: global/local lexical/stem/argument overlaps...
- ▶ For each measure, we calculated:

## Challenge

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

## Improvement

Complexity(continuation writing) – Complexity(baseline writing)

- ▶ Linear regression models: improvement ~ challenge

- ▶ Challenge is able to account for the variance of improvement in 88% (491/558) of the complexity measures.
- ▶ R-squared range from 8% to 93%.
- ▶ Challenge explains most variance in improvement for:
  - ▶ Number of past-tense verb types (94%)
  - ▶ Mean token frequency of all words in BNC (89%)
  - ▶ SD of token frequency of adverbs in BNC (88%)
  - ▶ Number of coordinating conjunctions (85%)
- ▶ Measures showing little effect:
  - ▶ Mean freq. of adj. types in Subtlex Log10 word frequency (8.4%)
  - ▶ Number of passive sentences (8.5%)
  - ▶ Number of easy words from the NGSL list (8.8%)
  - ▶ Easy adverbs STTR from the BNC (8.9%)

# Lexical Soph.: Mean Token Frequency from BNC

## Introduction

Relating input to  
production using  
complexity analysis

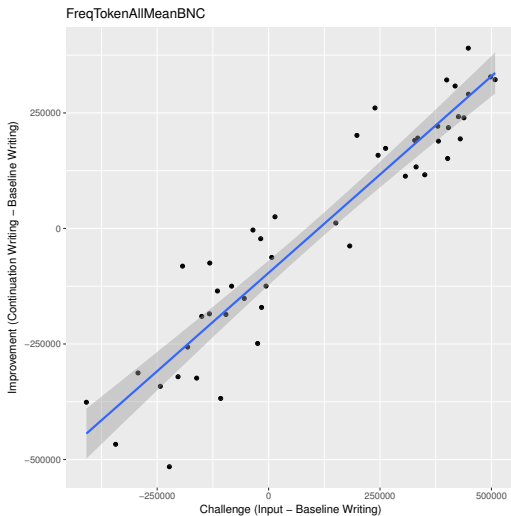
## Results

Findings

Application

Outlook

References



$$(r^2 = 89\%)$$

# Lexical Variation: Corrected TTR of Verbs

Alignment in  
Linguistic  
Complexity

Xiaobin Chen  
Detmar Meurers

Introduction

Relating input to  
production using  
complexity analysis

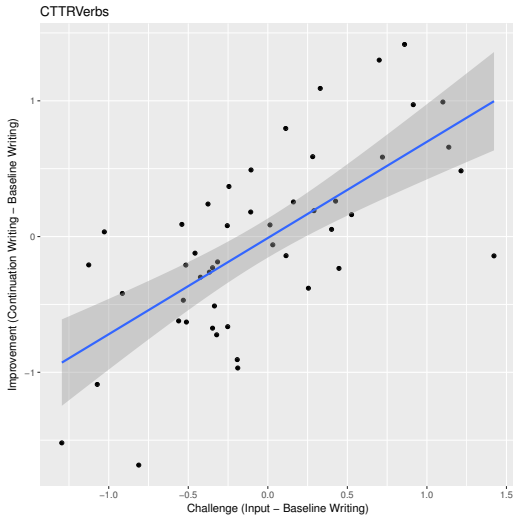
Results

Findings

Application

Outlook

References



$$(r^2 = 61\%)$$

EBERHARD KARLS  
UNIVERSITÄT  
TÜBINGEN



LEAD  
Graduate School & Research Network



# Syntactic Comp.: # Clauses per T-unit

## Introduction

Relating input to  
production using  
complexity analysis

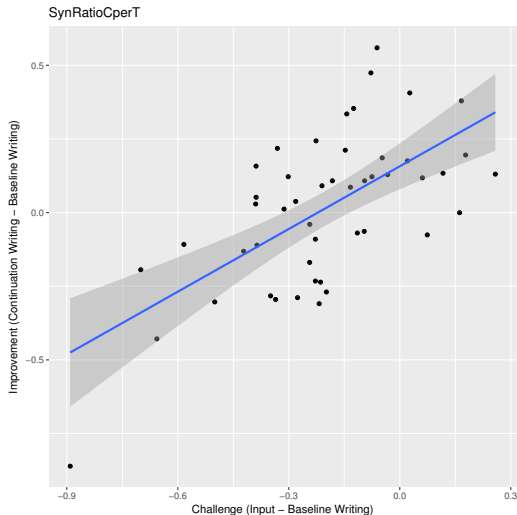
## Results

Findings

Application

Outlook

References



$$(r^2 = 68\%)$$

# Cohesion: Mean Global Lexical Overlap

## Introduction

Relating input to  
production using  
complexity analysis

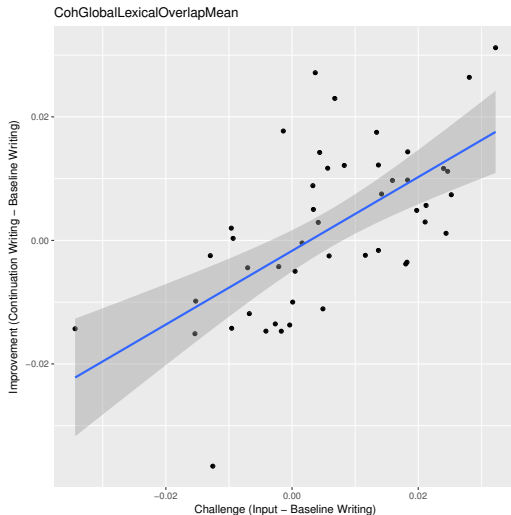
## Results

### Findings

### Application

### Outlook

### References



$$(r^2 = 47\%)$$

- ▶ Not only is alignment manifested as fewer errors in production, but also as alignment of the variedness and elaborateness between input and production.
- ▶ Negative alignment (i.e. aligning to less complex input) is also observable besides positive alignment.
- ▶ Alignment happens on multiple linguistic levels.

- ▶ SyB: Challenging learners in their individual ZPD (Chen & Meurers 2017)
- ▶ Pedagogic corpus as benchmark of complexity development
  - ▶ Designed to exhibit a linear development of complexity measures
  - ▶ Well-formed language easy to process with NLP tools
- ▶ How it works?
  - ▶ Analyze the complexity of learner production
  - ▶ Place learner complexity level on benchmark scale
  - ▶ Challenge learner with reading texts of higher complexity
  - ▶ Repeat the process to promote development of interlanguage
  - ▶ Realizes Krashen's  $i+1$

Introduction

Relating input to  
production using  
complexity analysis

Results

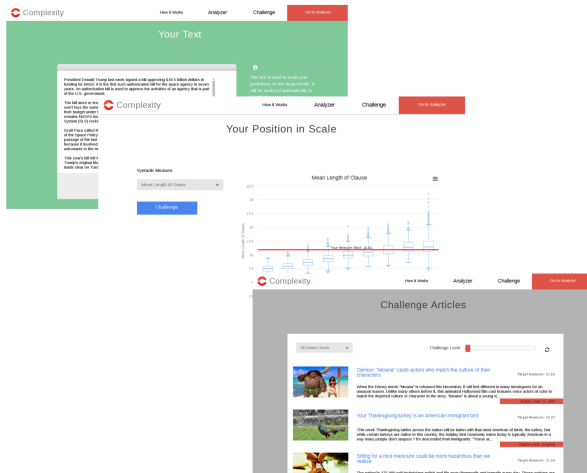
Findings

Application

Outlook

References

# Availability



The SyB system is available at: <http://complexityweb.org>

Alignment in  
Linguistic  
Complexity

Xiaobin Chen  
Detmar Meurers

Introduction

Relating input to  
production using  
complexity analysis

Results

Findings

Application

Outlook

References

# Ongoing Project

- ▶ Randomized control experiment: Complex Input Primed Writing (CIPW)
- ▶ Control for general complexity of input
- ▶ Treatment: varying challenge levels of one complexity measure (How big should the  $+1$  be?)
- ▶ Task: continuation writing
- ▶ Input texts: selected based on the complexity of learner production (the  $i$ )
- ▶ Purposes of study:
  - ▶ to figure out how much challenge is within the learner's ZPD (where does the developmental line level off?)
  - ▶ to investigate the interactional effects of challenge and proficiency on the complexity development of interlanguage
  - ▶ Probably also: interactional effects of challenge and individual differences

## Introduction

Relating input to  
production using  
complexity analysis

## Results

## Findings

## Application

## Outlook

## References

The screenshot shows a web browser window with the address bar displaying 'localhost:8081/mlat-instructions-en.html'. The page title is 'CIPW - Complex In...'. The main content area is titled 'Modearn Language Aptitude Test'. It contains the following text:

You are going to take a test to determine how easily and rapidly you learn a foreign language. Do your best, regardless of whether or not you think you are good at learning foreign languages.

Your first task is to **MEMORIZE** some Kurdish-English vocabulary. You will be given **TWO** minutes to study the vocabulary on the next page. At the end of the two minutes, you will complete some fill-in-the-blank exercises, which will help you memorize the vocabulary. After filling the blanks, continue studying if there is still time.

Go ahead now and start memorizing the vocabulary. You will be tested on this later.

Please click on the "Start" button to proceed.

A blue 'Start' button is visible below the text.

The left sidebar contains the following navigation menu:

- Information
- BACKGROUND
- Questionnaire
- Motivation
- Proficiency
- MEMORY TESTS
  - OSpan
  - CVMT
  - MLAT
- CIPW TASKS
  - Pre-test Writing
  - CIPW 3 / 10
  - Post-test Writing

- ▶ Linguistic complexity can be used to relate input to production, empirically implementing Krashen's  $i+1$  hypothesis.
- ▶ Linguistic complexity is a better account of alignment in language learning than the other methods.
- ▶ Research-informed ICALL systems are likely to benefit L2 learners.



# References

- Chen, X. & D. Meurers (2017). Challenging Learners in Their Individual Zone of Proximal Development Using Pedagogic Developmental Benchmarks of Syntactic Complexity. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017*. ACL, Linköping Electronic Conference Proceedings 134, pp. 8–17. URL <http://aclweb.org/anthology/W17-0302.pdf>.
- Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford, UK: Oxford University Press.
- Housen, A., F. Kuiken & I. Vedder (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken & I. Vedder (eds.), *Dimensions of L2 Performance and Proficiency*, John Benjamins, Language Learning & Language Teaching, pp. 1–20.
- Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Addison-Wesley Longman Ltd.
- Rescher, N. (1998). *Complexity: A philosophical overview*. London: Transaction Publishers.
- Vygotsky, L. S. (1986). *Thought and Language*. Cambridge, MA: MIT Press.
- Wang, C. & M. Wang (2015). Effect of alignment on L2 written production. *Applied Linguistics* 36(5).