Automating Measures of L2 Syntactic Complexity

Xiaofei Lu

Department of Applied Linguistics

Pennsylvania State University

Outline

- L2 syntactic complexity
- Task and motivation
- Measures to be automated
- The approach
- Evaluation
- Conclusion

L2 Syntactic complexity

- What is syntactic complexity
  - The range and degree of sophistication of forms that surface in language production (Ortega 2003)
- Why measure L2 syntactic complexity
  - Evaluating learner's developmental level in L2
    - What (advanced learners) *can* do (vs. errors)
  - Assessing cross-proficiency differences
  - Assessing effect of pedagogical intervention

Task and motivation

- Task: to automate L2 syntactic complexity analysis using 14 different metrics
  - Targeting college-level English learners
- Motivation
  - Limit on sample size and number of measures
  - Easy comparison of metrics
    - Previous results mixed
    - Research synthesis problematic
  - Informing automatic essay scoring

Measures to be automated

- All proposed in one or more SLA studies
- Measures reviewed in 2 research syntheses
  - Wolfe-Quintero et al. (1998)
  - Ortega (2003)
- Selection criterion
  - At least one previous study showed at least weak correlation with or effect for proficiency

Measures to be automated (cont.)

- Length of production
  1. Mean length of clause (MLC)
  2. Mean length of sentence (MLS)
  3. Mean length of T-unit (MLT)
- Sentence complexity
  4. Mean number of clauses per sentence (C/S)

Measures to be automated (cont.)

- Subordination
  5. Mean number of clauses per T-unit (C/T)
  6. Mean number of complex T-units per T-unit (CT/T)
  7. Mean number of dependent clauses per clause (DC/C)
  8. Mean number of dependent clauses per T-unit (DC/T)

Measures to be automated (cont.)

- Coordination
  9. Mean number of coordinate phrases per clause (CP/C)
  10. Mean number of coordinate phrases per T-unit (CP/T)
  11. Mean number of T-units per sentence (T/S)

Measures to be automated (cont.)

- Particular grammatical structures
  12. Mean number of complex nominals per clause (CN/C)
  13. Mean number of complex nominals per T-unit (CN/T)
  14. Mean number of verb phrases per T-unit (VP/T)

The approach

- Input: plain text
- Step 1: Syntactic parsing using Stanford parser
- Step 2: Counting occurrences of the following
  - Words, sentences, clauses, dependent clauses
  - T-units, complex T-units, coordinate phrases
  - Complex nominals, verb phrases
- Step 3: Computing ratios for the 14 measures
- Output: 14 syntactic complexity indices

The approach (cont.)

- Word: all non-punctuation tokens
- Other units: Tregex (Levy & Andrew, 2006)
  - Define the units linguistically
  - Formulate patterns that match the unit definitions
  - Query the parse trees with the Tregex patterns
  - Retrieve/count (sub)trees matching each pattern

Examples of defined patterns

Sentence: as punctuated by user, including fragments

'ROOT'

T-unit: a main clause plus any subordinate clauses (Hunt 1965); includes sentence fragments (Tapia 1993)

'S|SBARQ|SINV|SQ > ROOT | [$-- S|SBARQ|SINV|SQ !>> SBAR|VP]'

'FRAG > ROOT'

Examples of defined patterns (cont.)

Clause: subject + finite verb (Polio 1997); S fragments with no overt verb (Bardovi-Harlig & Bofman 1989)

'S|SINV|SQ < (VP <# MD|VBP|VBZ|VBD)'

'FRAG > ROOT !<< VP'

Dependent clause: adverbial, adjectival or nominal clause

'SBAR < (S|SINV|SQ < (VP <# MD|VBP|VBZ|VBD))'

Evaluation

- Comparison with manual analysis
  - 10 essays from WECCL (Wen et al. 2005), average length 315 words
  - Two annotators counted the number of occurrences of the various units
- Parse and pattern quality affects performance

Raw count correlations

	S	C	DC	T	CT	CP	CN	VP
A1-A2	1.000	.991	.967	.996	.896	.973	.915	.971
Sys-A1	1.000	.957	.929	.993	.860	.866	**.859**	.943
Sys-A2	1.000	.965	.946	.996	.952	.903	.954	.971

Complexity score correlations

Measure	A1-A2	Sys-A1	Sys-A2	Measure	A1-A2	Sys-A1	Sys-A2
MLC	.985	.946	.967	DC/T	.981	.956	.968
MLS	1.000	1.000	1.000	CP/C	.964	.874	.943
MLT	.998	.993	.994	CP/T	.965	**.842**	.899
C/S	.978	.944	.957	T/S	.969	.930	.942
C/T	.978	.984	.963	CN/C	.948	.886	.959
CT/T	.912	.902	.967	CN/T	.957	.905	.969
DC/C	.954	.854	.846	VP/T	.958	.877	.944

The effect of errors?

- It’s the basic require of civilized people.

```
(ROOT
  (S
    (NP (PRP It))
    (VP (VBZ 's)
      (NP
        (NP (DT the) (JJ basic) (NN require))
        (PP (TO to)
          (NP (DT the) (JJ civilized) (NNS people)))))
    (. .)))
```

The effect of errors?

- Honesty [[makes people [tell the truth]] and [works without [cheating, which could make the society more finer.]]]

  S: 1-1

  T: 1-1, CT: 1-1

  C: 2-2, DC: 1-1

  CP: 1-1

  CN: 1-0

  VP: 3-3

Summary

- A tool for automatic analysis for syntactic complexity using 14 measures
  - Works well with written samples produced by college-level English learners
- More rigorous evaluation planned
  - Unit by unit evaluation

An application of the tool

- An evaluation of these measures as indices of college-level ESL writers' language proficiency
  - Data: 3,554 essays from WECCL
  - Proficiency operationalized using school levels

An application of the tool (cont.)

- Research questions answered
  - Impact of sampling condition on the measures
  - Which measures significantly differentiate between-proficiency levels?
  - What magnitudes are required for between-proficiency differences to be significant?
  - How do the measures relate to each other?
  - What are the patterns of development associated with each measure?