Automatic Measurement of Syntactic Complexity in SLA

Xiaofei Lu
Department of Applied Linguistics
Pennsylvania State University
Outline

- Syntactic complexity
- Syntactic complexity metrics in SLA
- Automating complexity analysis
- Corpus-based metrics evaluation
Syntactic complexity

- What is syntactic complexity
  - The range of forms that surface in language production and the degree of sophistication of such forms (Ortega 2003:492)
Why measure syntactic complexity?

○ Theoretical linguistics
  ● Cross-linguistic comparisons of relative complexity of grammatical constructions

○ Psycholinguistics
  ● Developmental trends in child grammar acquisition
  ● Ordering experimental stimuli by complexity
  ● Low complexity in childhood vs. symptoms of Alzheimer’s disease (Kemper et al. 2001)
Why measure syntactic complexity?

- Second language acquisition
  - Evaluating effect of pedagogical intervention on the development of grammar or writing ability
  - Assessing differences in L2 texts written by learners across proficiency levels
Syntactic complexity metrics

- Metrics differ for L1 and L2
- Review of complexity metrics for L1
  - Cheung & Kemper (1992)
- Review of complexity metrics for L2
  - Wolfe-Quintero et al. (1998)
  - Ortega (2003)
Syntactic complexity metrics for L1

- **Sentence length**
  - Mean clauses or length per utterance (MCU/MLU)

- **Type of grammatical construction**
  - Developmental Level (D-Level)
  - Developmental Sentence Scoring (DSS)
  - Index of Productive Syntax (IPSyn)

- **Processing demand**
  - Frazier’s node count
  - Ygnve depth
Syntactic complexity measures in SLA

- **Length of production**
  1. Mean length of sentence (MLS)
  2. Mean length of T-unit (MLTU)
  3. Mean length of clause (MLC)

- **General complexity measures**
  4. Mean number of clauses per T-unit (C/T)
  5. Mean number of clauses per sentence (C/S)
Syntactic complexity measures in SLA

- Amount of coordination
  6. Mean number of T-units per sentence (T/S)
  7. Mean number of coordinate clauses per T-unit (CC/T)
  8. Mean number of coordinate phrases per T-unit (CP/T)
Syntactic complexity measures in SLA

- Dependent clause measures
  9. Mean number of adverbial clauses per clause (AC/C)
  10. Mean number of adverbial clauses per T-unit (AC/T)
  11. Mean number of dependent infinitives per T-unit (DI/T)
Syntactic complexity measures in SLA

- Presence of particular grammatical structures
  12. Mean number of complex nominals per T-unit (CN/T)
  13. Mean number of passives per T-unit (P/T)
  14. Mean number of passives per clause (P/C)
  15. Mean number of passives per sentence (P/S)
Automating complexity analysis

- What is the task?
  - Automatic complexity analysis of samples of learner language using all 15 metrics for SLA

- Why bother?
  - Manual analysis laborious and costly
  - Sample size affects reliability of results

- System structure
  - Preprocessor
  - Syntactic complexity analyzer
Preprocessor

- Sentence and word segmentation
- Part-of-speech tagging
- Syntactic parsing
Syntactic complexity analyzer

- Detecting clause and T-unit boundaries
- Identifying the following constructions
  - Coordinate clauses and phrases
  - Adverbial clauses
  - Dependent infinitives
  - Complex nominals
  - Passives
Syntactic complexity analyzer

- Counting the number of
  - Words, clauses, T-units, and sentences
  - Each of the relevant syntactic constructions

- Computing the ratios for each measure
Challenges with learner language

- Imperfect language input
  - Spelling mistakes
  - Incomplete sentences
  - Ungrammatical sentences

- Addressing potential problems for the parser
  - Lessons from Lu (2008)
  - Using shallow parsing
Lessons learned from Lu (2008)

- Implementation of the D-Level scale
  - Assigns sentence to 8 levels based on structure
  - Works with spoken child language acquisition data
- Not all errors cause a problem for the parser
  - Agreement and tense errors
  - Word and collocation errors
- Not all parsing errors affect analysis
Using shallow parsing

- May use information about verbs to
  - Approximate number of clauses and T-units
  - Identify passives
Corpus-based metrics evaluation

○ What is the task?
  ● Evaluating adequacy of the 15 metrics for differentiating levels of syntactic acquisition

○ Why bother?
  ● The search for the “best” metrics (Wolfe-Quintero et al. 1998; Ortega 2003)
  ● The first corpus-based study in such a search
Data and results

- Data for training and evaluation
  - International corpus of learner English (Granger, Dagneaux & Meunier 2002)
  - Spoken and Written English Corpus of Chinese Learners (Wen, Wang & Liang 2005)

- Results
  - Will be released in the summer...
Acknowledgements

- Thanks to the following people for useful discussion on this project
  - James Lantolf
  - Diane Larson-Freeman