Computational Approaches to Second Language Acquisition

Dora Alexopoulou
(with Jeroen Geertzen & Anna Korhonen)
Dept of Theoretical and Applied Linguistics[1]

Language Sciences in the 21st Century:
the interdisciplinary challenge
Cambridge, October 2013

[1] Research supported by EF Education First and the Isaac Newton Trust

Aims of this talk

Relevenance of Computational Linguistics for:

- SLA theory: grammar induction research relevant for modeling L2 grammar acquisition.
- Big educational learner data: Natural Language Technology for learner language to unlock information in big data for SLA research.

Questions linguists ask

- What is linguistic knowledge like?
- How do humans acquire such knowledge?
- How can we account for linguistic diversity, and the common patterns in linguistic systems crosslinguistically?

Chomsky-Generative Grammar

- Natural language syntax is a context free grammar; syntactic categories can be captured by (binary) formal features (Chomsky 1955,1957,1970).
- *A priori* knowledge of *Universal Grammar*, a set of principles constraining natural language (universals) and a set of parameters defining the range of variation across languages (linguistic diversity).

Usage-based theory of acquisition and Construction Grammar

- Grammatical knowledge entirely derivative from general socio-cognitive capacities of humans: reading of communicative intentions and ability to *refer* using symbolic representations; syntactic acquisition supported by generalisation mechanisms (functional analogy, distributional categorisation)—(Tomasello 2003).
- Construction Grammar: referential form-meaning mappings develop *through usage* to syntactic categories with combinatorial possibilities. From item-specific to more abstract generalisations (Goldberg, 2006).

Developing Grammars

- Gemerative grammar: results on developmental stages but not on *transition* from one developmental stage to the next (Young 2011).
- Usage-based theory: emphasis on the general social and cognitive abilities of humans, detailed work on early constructions, but remains an informal theory (Bod 2009, but see Sag et.al 2012).
- Desideratum: a formal theory of *developing grammars* linking the growth of grammatical knowledge with the acquisitional mechanisms that enable it.

From probable to possible grammars

- Grammar induction for real life applications (Clark and Lappin 2007).
- Grammar induction algorithms are relevant for modeling human language acquisition.
- Probabilistic nature of the acquisition algorithms and stochastic grammars.
- Focus from what is *possible* in a grammar to what is *probable* (Newmeyer 2005, Manning 2003).
- What is possible can be derived from what is probable over the course of acquisition.
- But insightful models of acquisition necessitate the inclusion of meaning and contextual information.

Second Language Acquisition

SLA questions:

- SLA: enormous individual and contextual variation in learning (age of onset, educational background, formal instruction vs. immersion etc).
- But L2 learners know *what* they want to say, they need to learn *how*. Focus on grammar acquisition.
- L1 effects as varying structural biases at initial stages of acquisition.

SLA data:

- Large amounts of machine readable data that can support computational modeling.
- Constrained semantic/discourse descriptions of scripts.

An Interdisciplinary Challenge for SLA

A formal theory of *L2 developing grammars* linking represenations of grammatical knowledge with the acquisitional mechanisms enabling such knowledge.

- Computational modeling key to such a theory.
- Incorporate meaning to the learning models.

Big SLA educational data

- Assessment and Educational Institutions with a global reach, often exploiting online platforms create huge amounts of educational data which can lead to big data resources for research in SLA and Education: e.g. Cambridge Learner Corpus, EF Cambridge Open Language Database (EFCAMDAT).
- Challenges: size and unpredictability of data (e.g. task effects, incomplete patterns for individual learners etc.).

NLP technology for SLA research

1. Annotations of learner language rich enough for SLA research.
2. Classification of messy data.
3. Data-pattern extraction to lead to new ’observations’.

Annotating learner language

- Automatic annotation for errors, parts of speech and grammatical relations (De Feliche 2008, Boyd-Meurers 20011, Andersen 2010, Kochmar et. al. 2012, Geertzen et. al 2012, Rosen et. al. 2013).
- But annotations not rich enough for SLA research.

(1) I know, **what he likes reading books.** Therefore we giving him a book. I know, **what he likes eating chocolat.** Therefore we giving him box of chocolates. And I know, **what he likes flowers,** therefore we need to buy a bouquet of flowers. (Level 3, Russian)

1. What is the correct analysis of examples like (1) (note, Russian is a zero copula language)?
2. Is this a "systematic" pattern/structure of a particular stage of learner language of a particular learner group?

Annotating learner language

- Automatic annotation for errors, parts of speech and grammatical relations (De Feliche 2008, Boyd-Meurers 20011, Andersen 2010, Kochmar et. al. 2012, Geertzen et. al 2012, Rosen et. al. 2013).
- But annotations not rich enough for SLA research.

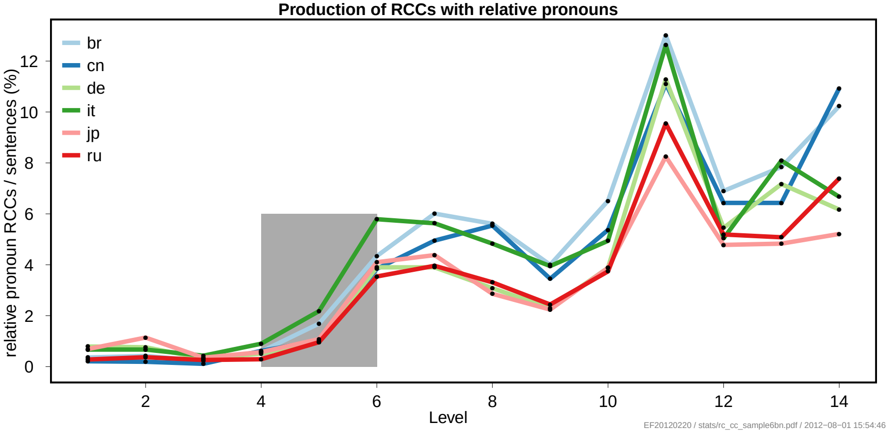
(1) I know, **what he likes reading books.** Therefore we giving him a book. I know, **what he likes eating chocolat.** Therefore we giving him box of chocolates. And I know, **what he likes flowers,** therefore we need to buy a bouquet of flowers. (Level 3, Russian)

1. What is the correct analysis of examples like (1) (note, Russian is a zero copula language)?
2. Is this a "systematic" pattern/structure of a particular stage of learner language of a particular learner group?

Understanding messy data: mapping relative clauses

Relative clause: *In the end, the person who has the most points is the winner*.

Cross-sectional RC production by "L1"



EF20120220 / stats/rc_cc_sample6bn.pdf / 2012−08−01 15:54:46

(2) a. I had to married a aweful man that i don’t love for some reasons.

b. **For those of you that don’t know me**.My name is Liji Yuan......Here is an interesting fact.Do you know that the more companies are interesting in this products?

How can we distinguish formulaic relatives, from relatives elicited by a specific task, from productive ones?

Picking formulaic and task related relatives

Measure combining frequency of use by learners and internal coherence:

(3) a. ...those of you that don't know me ... (Level 1, Unit 5).
b. ... let me tell you what I did... (Level 4).
c. ... for each pin that is knocked down (Level 7, Unit 1).

Making new observations

- Data-driven NLP technology for analysis of data and patterns to provide novel observations about the data.
- Poster session!
  Yannakoudakis H., T. Briscoe, T. Alexopoulou, Automating L2 Acquisition Research: an interdisciplinary perspective.

Conclusion

- NLP technology vital for exploitating Big SLA/educational data.
- Computational modeling a crucial component for a formal theory of L2 acquisition.