Automating L2 acquisition research: an interdisciplinary perspective

***Helen Yannakoudakis*[1]*, Ted Briscoe*[1]*, Dora Alexopoulou*[2]**

[1]*Computer Laboratory, University of Cambridge;* [2]*DTAL, University of Cambridge.*

UNIVERSITY OF CAMBRIDGE

Introduction

L2 acquisition research attempts to explain how the learner acquires a second language, and make inferences about the internal representations of L2 knowledge in the mind of the learner at different stages of learning. L2 research dimensions include:

- Why do some learners do better than others?
- Is there a native language (L1) impact and transfer effects?
- What types of error emerge at different learning stages?
- What developmental sequences do learners pass through?
- What L2 features are acquired at different learning stages?

Common methodologies in L2 acquisition research involve theory-driven approaches for formulating hypotheses on learner grammars, typically based on linguistic intuition and the extant literature on learner English.

Advantage: allow us to identify learner-language properties that are well understood and can inform learning theory.

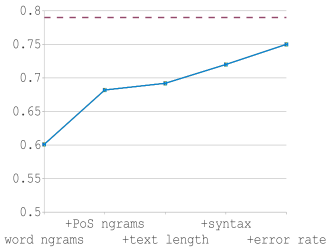
Disadvantage: may emphasise self-evident hypotheses and overlook properties about learner grammars that may not have been discussed in the linguistic literature.

Approach: we propose a new methodological technique to L2 research and use data-driven methods in tandem with visualisation techniques to shed light on understanding the linguistic abilities that characterise different levels of attainment and, more generally, developmental aspects of learner grammars.

Advantages: (i) automated hypothesis formation, (ii) more empirical approach, (iii) exploration of a much wider hypothesis space, (iv) quantitatively very powerful, (v) effective exploration through visualisation, (vi) useful adjunct to theory-driven approaches.

Methodology

- Use discriminative machine learning methods to automatically identify linguistic features that are predictive of a learner's level of attainment (Fig. 2).
- Apply coordinated graph visualisation techniques on discriminative features and use them as a tool for hypothesis generation (Fig. 3).



Feature	Example
AP/a1 (+)	*(very) clever(ly)*
NP/det_n1 (+)	*the movie*
VM_RR (+)	*could clearly*
,_because (−)	*, because of*
necessary (+)	*it is necessary that*
the_people (−)	*the people are clever*
P1/p_s (+)	*after we left*
how_to (−)	(learn) how to
NN2_VVG (+)	*children smiling*

Figure 2: *Highly predictive discriminative features.*

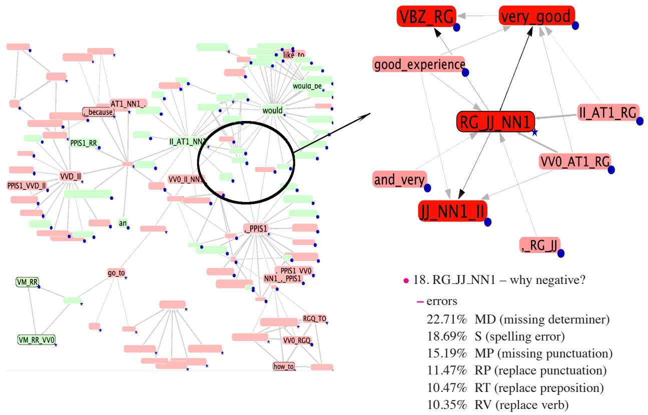


Figure 3: *Linking together the statistical and visual components of a 'feature network' to facilitate hypothesis formation.*

Results

(1a) *Unix* $\overbrace{\textbf{\textit{is very powerful system}}}^{\textit{VBZ RG JJ NN1}}$ *but there is one thing against it.*

(1b) *I think it's* ***very good idea to*** *spending vacation together.*

Emerging picture: there is a link between richer nominal structures that include more than one modifier and article omission. Two questions arise: (i) why these richer nominals should associate with article omission and (ii) why only singular nouns are implicated in this feature.

Language	sentences% RG_JJ_NN1	sentences% VBZ_RG_JJ	MD:doc RG_JJ_NN1	MD:doc VBZ_RG_JJ
all	23.0	15.6	2.75	2.73
Turkish	45.2	29.0	5.81	5.82
Japanese	44.4	22.3	4.48	3.98
Korean	46.7	35.0	5.48	5.31
Russian	46.7	23.4	5.42	4.59
Chinese	23.4	13.5	3.58	3.25
French	6.9	6.7	1.32	1.49
German	2.1	3.0	0.91	0.92
Spanish	10.0	9.6	1.18	1.35
Greek	15.5	12.9	1.60	1.70

Table 1: *Feature relations with MD errors for different L1s, where sentences% shows the proportion of sentences containing the feature that also contain a MD, and MD:doc shows the ratio of MD errors per script. We can see that there is a sharp contrast between L1s with articles (French, German, Spanish and Greek) and those without (Turkish, Japanese, Korean, Russian, Chinese).*

Task set	System	Time (min) Mean	Time (min) SD	Mouse events Mean	Mouse events SD	Task accuracy%
Set A	Search tool	6.38	3.48	35.62	20.20	69
Set A	Visualiser	1.06	0.54	5.12	3.30	94
Set B	Search tool	2.44	1.93	14.06	12.12	62
Set B	Visualiser	1.62	1.26	5.06	3.43	75

Table 2: *Mean and standard deviation (SD) for task completion time and mouse event (click) counts, as well as task accuracy when participants used the Visualiser (our system) or a standard search tool to complete task set A, focusing on hypothesis generation through discriminative features, and task set B, focusing on learner corpus statistics.*

System	Usefulness Mean	Usefulness SD	Ease of use Mean	Ease of use SD	Ease of learning Mean	Ease of learning SD	Satisfaction Mean	Satisfaction SD
Search tool	3.62	0.34	3.95	0.78	4.94	0.72	3.71	1.11
Visualiser	5.16	0.89	5.23	0.50	5.50	0.41	5.50	0.36

Table 3: *Mean and standard deviation (SD) for subjective satisfaction scores, measured using the USE questionnaire (Lund, 2001), when participants used the Visualiser (our system) or a standard search tool to complete all tasks. Satisfaction scores range from 1 (strongly disagree) to 7 (strongly agree).*

Conclusions

- Demonstrated the usefulness of machine learning in tandem with visualisation techniques towards automating hypothesis formation about learner grammars.
- Nominals with complex adjectival phrases appear particularly susceptible to article omission errors by learners of English with L1s lacking articles.

Acknowledgements: Cambridge English Language Assessment and EF Education First. Thanks to Ø. Andersen, F. Barker and A. Salamoura.