Formulaic vs productive language in big learner data

J. Geertzen & T. Alexopoulou & A. Korhonen & D. Meurers

Education First

UNIVERSITY OF CAMBRIDGE
Department of Theoretical and Applied Linguistics

1. Introduction

- Online schools for learning English collect large amounts of learner data as part of their operations, which have recently become available to research in second language acquisition (SLA) [2].
- Accurately characterising language development requires the identification of language pieces as being productive, formulaic, under- or over-represented.
- But annotating large amounts of data by hand is prohibitively expensive and motivates the use of natural language processing techniques.
- **How could productive language be distinguished from formulaic language or language overrepresented because of input/task effects?**
- The relative clause was taken as a study case. E.g.: "The guys [who you met today] are not cool."

2. Learner data

- We used EFCamDat [2], a novel learner corpus that contains essays submitted to *Englishtown*, an online school of EF Education First.
- Englishtown offers 16 proficiency levels aligned with common standards such as TOEFL and IELTS. Each level contains 8 lessons with receptive and productive tasks. EFCamDat comprises scripts of the writing task that ends each lesson.
- The first release contains 551,036 scripts produced by 84,864 learners from 172 nationalities.
- Grammatical dependency relations were obtained with the Stanford and the C&C parser [1]. Evaluation on 1,000 sentences showed that relative clause modifiers could be recovered relatively well, at an F-score (combining precision and recall) of ±92%.

3. 'Fixed' expressions

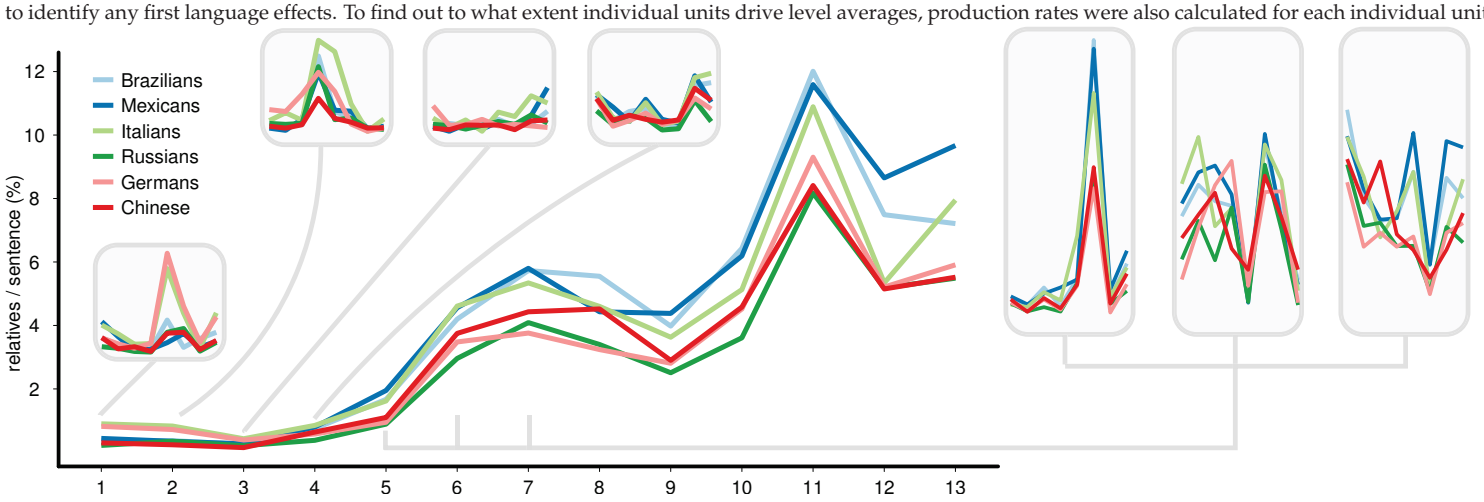
- To find formulaic or lifted expressions, sentences were described by word n-grams of various lengths, to identify those that:
  1. are produced often (F)
  2. are produced by quite some learners (S)
  3. have considerable word length
  4. consist of words that as a sequence occur often relative to how often each word occurs on its own. Pointwise Mutual Information (PMI) is a statistic that measures such association. For a bigram of word x and word y:

$$PMI(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)}$$

- Ranking and thresholding n-grams on a combination of above chacterstics picks language that stands out.

4. Relative clause production

- Relatives introduced by complementizers (e.g. *that*) or relative pronouns (e.g. *which*, *who*) were tracked and production rates calculated for the top six nationalities, to identify any first language effects. To find out to what extent individual units drive level averages, production rates were also calculated for each individual unit.



- Relative clause use seems to pick up from level 4 to 6 and then increases steadily with proficiency with a sharp peak at level 11. Learners of different nationalities follow a similar longitudinal path, but Brazilians and Mexicans produce more relatives than others.
- There is considerable fluctuation between adjacent lessons, which is partly caused by differences in the nature of the writings task or topic.
- Production rates may also be affected by learners using formulaic language, or 'lifting' expressions from input.

5. Formula's & lifted expressions

- Top-ranking n-grams of Brazilian students produced by at least 10% of learners:

Level	*n-gram*	*F*	*S*	*PMI*	*S*PMI*
1.5	**for those of you that do not know me my name is**	140	51	54	2730
4.7	**let me tell you what I did**	114	52	26	1363
5.8	there are things that we should remember to do	14	9	39	371
6.3	anyone who does not follow the dress code will lose their job	31	14	65	926
6.4	a job that allows me to use my	73	37	34	1289
6.6	here is a plan that might work for you	45	19	44	810
7.1	**for each pin that is knocked down**	554	30	29	866
8.3	things that I would like to do	27	13	28	347
8.6	will be assigned an instructor who will	8	12	33	411
10.3	that I will have to pay off the loan	28	13	53	698
11.7	allows the company to refuse to pay me for something that is not	10	16	64	998
12.3	the sand painting that you	21	28	17	473

- The two items with the highest score involve frozen expressions, possibly 'lifted' from the input
- By contrast, item 7.1, produced by around a third of learners, is quite likely a productive piece of language overrepresented in the learners' use because of a particular writing task.
- Items that are produced by the most learners tend to occur at lower proficiency levels, in line with the fact that as learners' proficiency advances, they rely less and less on formulaic language [3].

6. Conclusions

- Data consistency is an important methodological aspect of any language study, but is particularly important for big learner data obtained outside a lab environment such as featured by EFCamDat.
- We have shown, using relative clauses as a study case, how selected measures could be used effectively to provide a way to distinguish productive from formulaic language, or language overrepresented because of input or task effects.
- This aids SLA research on big learner data as it enables more accurate analyses of longitudinal changes in productive language use

References

[1] Clark, S., Curran, J.R. (2007). *Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Model* CL, 33, 495–552.

[2] Geertzen, J., Alexopoulou, T., Korhonen, A. (2013). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat) In *Proc. of the 31st SLRF*.

[3] Myles, F., (1995). *Interaction between linguistic theory and language processing in SLA* Second Language Research, 11, 235–266.