Contextually modulated syntactic variability in child-directed speech

Aaron Steven White
University of Maryland

IASCL 2014
Amsterdam, The Netherlands

1. Introduction

1.1 Grammar learning and syntax-based word learning

Common (implicit) assumption
CDS homogenous with respect to context

1.2 Syntactic complexity across context

Lower in CDS than adult speech (Buttery & Korhonen 2005)
Conclusion: not enough info for syntactic bootstrapping? (Gleitman 1990)

1.3 Intuition

Child-Ambient Speech (CAS) mixture of CDS+adult-speech

1.4 This study

Hypothesis: Complexity in CAS contexts non-homongenous
Example: play contexts less complex than dinner contexts for some verbs (Ely et al. 2001)
Experiment: syntactic complexity of dinner contexts > play
Conclusion: must account for contextual variability in models

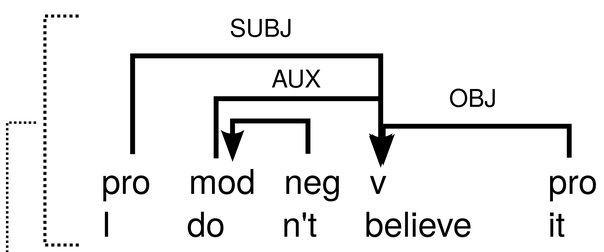
2. Dataset

2.1 Base corpus

Gleason (Masur & Gleason 1980)
24 children (age: 2;1-5;2, sex: 12 females)
One play session with mother; with father
Dinner session with mother and father

2.2 Annotation

GRASP dependency parses (Sagae et al. 2007)



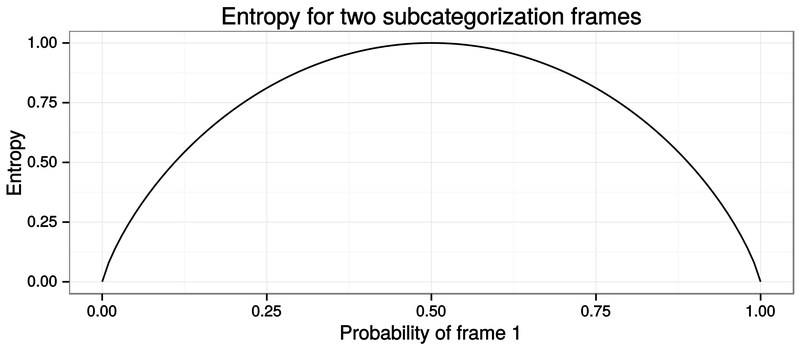
2.3 Extraction

For every verb token produced by MOT or FAT
a. Extract relation sequence
b. Stem verb
c. Remove auxiliary and adverb relations

Verb	Subcorpus	Participant	Frame
believe	dinner	david	SUBJ_OBJ

2.4 Complexity measure

Estimate each verb's syntactic distribution by child+context from dataset then calculate entropy of each syntactic distribution



2.5 Bootstrapping complexity statistics

Problem #1: Larger dataset will have higher average complexity
Solution: Match corpus size by subsampling larger subcorpus to size of smaller by-child

Problem #2: Power law distributions of frames
Solution: Bootstrap by-verb entropy from subsampled datasets

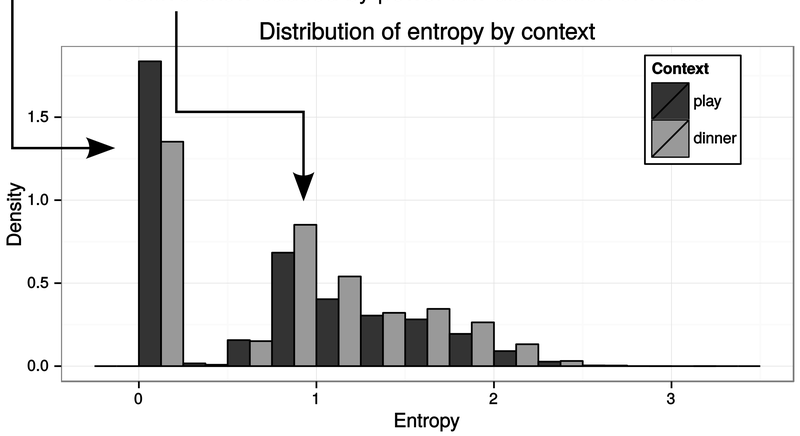
3. Results

3.1 Mean entropy

Context	Estimate	95% CI
dinner	0.812	[0.798, 0.844]
play	0.661	[0.686, 0.643]

3.2 Problems with raw means

Zero inflation caused by verbs that occur only once
Positive skew caused by power law distribution of verbs



3.4 Solution

Two component mixed model
1. Logistic component for zero-inflation
2. Inverse-gamma component for skew (best fit of gaussian or gamma with inv- or log-link)

Fixed effects for context and log(freq. of verb)
Random intercepts for child and verb

3.5 Logistic model

Term	Estimate	95% CI
Intercept (play)	-3.764	[-3.988, -3.436]
dinner	0.793	[0.522, 0.995]
log(freq)	3.364	[3.188, 3.619]

3.6 Inverse-gamma model

Term	Estimate	95% CI
Intercept (play)	1.205	[1.185, 1.227]
dinner	-0.026	[-0.038, -0.014]
log(freq)	-0.139	[-0.151, -0.127]

3.7 Summary

Logistic: dinner less zero complexity verbs controlling for freq
Gamma: dinner more complex in non-zero complexity verbs

4. Conclusion

4.1 Syntactic complexity modulated by context

Dinner contexts higher syntactic complexity than play context
Conclusion: must account for context variability in learning models

4.2 Future Directions

a. Mixtures of different adult genres in different CAS contexts?
Example: in the kitchen, at the bank, at the grocery store
b. Relationship between entropy and informativity in behavior?

Selected References

Buttery, P. and Korhonen, A. (2005) Large-scale analysis of verb subcategorization dierences between child directed speech and adult speech.

Ely, R., Gleason, J. Berko, MacGibbon, A., & Zaretsky, E. (2001). Attention to Language: Lessons Learned at the Dinner Table. Social Development, 10, 3, 355-373.

Gleitman, Lila. The structural sources of verb meanings. Language acquisition 1.1 (1990): 3-55.

Masur, E., & Gleason, J. B. (1980). Parent–child interaction and the acquisition of lexical information during play. Developmental Psychology, 16, 404–409.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B. and Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts.

Research supported by NSF BCS grant (1124338) and NSF DGE IGERT grant (0801465)