# Quantifying the Relationship Between Child and Caregiver Speech Using Generalized Estimating Equations: The Case of *only*

**Lindsay Hracs**
University of Calgary
lindsay.hracs@ucalgary.ca

## 1 Introduction

Research in first language acquisition often focuses on examining specific characteristics of child-directed speech (CDS) and child-produced speech (CPS). However, fewer studies investigate the direct connection between CDS and CPS, which is useful in determining what triggers changes in a child's grammar over the course of development. Using corpus analysis and quantitative modelling, this study fills a gap in the literature by determining if there is an explicit relationship between CDS and CPS relative to the acquisition of *only*.

Modelling data using Generalized Linear Models (GLMs) requires meeting the key assumption of independence among outcome measurements which is not possible when repeated measurements over time are being analyzed. Instead, Generalized Estimating Equations (GEEs) (originally presented in Liang and Zeger, 1986) are a particularly useful tool when examining longitudinal data as GEE models take the correlation between data points into consideration.

## 2 The Learning Problem

*Only* is an example of a focus sensitive particle. Focus sensitive particles are semantic operators that take scope over a constituent resulting in that constituent being construed as the focus, i.e. the information that a speaker wants a hearer to attend to (Erteschik-Shir, 1997). Consider (1).

(1) a. **Only** [Dale]$_{\text{FOCUS}}$ drinks coffee.
b. Dale **only** [drinks coffee]$_{\text{FOCUS}}$.
c. Dale drinks **only** [coffee]$_{\text{FOCUS}}$.
d. Dale drinks [coffee]$_{\text{FOCUS}}$ **only**.

The positional variability of *only* results in different interpretations. Given that *only* is situated at a grammatical interface, previous work (Crain et al., 1994; Kim, 2011; Paterson et al., 2006; among others) shows variability in the factors used to explain the complexity of this learning problem. Furthermore, these studies do not discuss the role of input, i.e. the linguistic stimuli that is analyzed by learners based on the current state of their grammars. Focusing on the distributional properties of *only* shown in (1), this study seeks to determine if the frequency of occurrence of *only* in CDS predicts the frequency of occurrence of *only* in CPS and whether distributional properties of *only* in CDS and CPS change over time.

## 3 The Dataset

This corpus study is part of a larger study designed to model the representational changes corresponding to the acquisition of *only* by English-speaking children. The dataset outlined here will be used as input to the learning model, however, the potential dynamic properties of CDS warrant an input analysis before being used to train learning algorithms.

A longitudinal dataset was constructed from the North American English corpora from the CHILDES database (MacWhinney, 2000). The dataset includes a total of 3,040 CHAT files from 511 different child-caregiver dyads, with child ages ranging from 0;3-9;9. The number of data points from each dyad range from 1 to 284. To avoid criticism that if the files are not matched for speaker-type there can be no measurable relationship, only files that contain both CDS and CPS were included. The overall observed frequency of *only* in CDS = 1,788 tokens and CPS = 920 tokens. Since each file had a different word count for CDS and CPS, relative frequencies (normalized to 1,000) were calculated. The overall relative frequency of *only* in CDS = 0.409 and CPS = 0.400. Distributional information, i.e. whether *only* was



*Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 381-382.
New York City, New York, January 3-6, 2019

in pre-subject, pre-verb, or pre-object position (see (1a-c) above), was extracted. Note that this study differs from existing work in that it investigates the development of *only* from a naturalistic and longitudinal perspective.

## 4 Analysis

Data cannot be considered independent when superordinate structures or repeated measurements create a correlation across data points (Vagenas and Totsika, 2018). GEEs provide estimates of regression parameters and parameter variance under the assumption of time dependence (Liang and Zeger, 1986) and clustering variables, e.g. family units, (Vagenas and Totsika, 2018), making them useful in modelling the relationship between CDS and CPS over time. Moreover, GEEs are robust to unbalanced designs, meaning that having only a single data point for some dyads and up to 284 for others does not affect parameter estimation. To account for the correlation, a clustering variable in the form of a unique identifier assigned to each of the dyads was included in the model. Finally, it is important to note that the models used in the analysis of these data employ maximum likelihood optimization methods.

Results show that the frequency of occurrence of *only* in CDS is a significant predictor of the frequency of occurrence of *only* in CPS (B=0.264, SE=0.077, $\chi^2(1)$=11.9, p<.001). Furthermore, the frequency of *only* in CDS significantly increases during development (B=0.008, SE= 0.002, $\chi^2(1)$=28.1, p<.001). When broken down by position, the frequency of *only* in CDS significantly increases for pre-subject position (B=0.002, SE=0.000, $\chi^2(1)$=26.4, p<.001) and pre-verbal position (B=0.003, SE=0.001, $\chi^2(1)$=5.96, p=.015), but not for pre-object position (B=0.000, SE=0.000, $\chi^2(1)$=0.02, p=.87). Regarding CPS, the frequency of *only* significantly increases during development (B=0.021, SE=0.003, $\chi^2(1)$=57.9, p<.001). When broken down by position, the frequency of *only* in CPS significantly increases for pre-subject position (B=0.006, SE=0.001, $\chi^2(1)$=49.0, p<.001) and pre-verbal position (B=0.007, SE=0.001, $\chi^2(1)$=32.9, p<.001), but not for pre-object position (B=0.000, SE= 0.000, $\chi^2(1)$=3.63, p=.057).

## 5 Discussion

The correlated and unbalanced nature of the dataset described above requires a model which is robust to violations of independence in order to determine if there is a relationship between CDS and CPS. GEEs, which are appropriate for modelling both longitudinal and clustered data, can be employed. Although it is not clear from the current analysis what is driving the change in frequency of *only*, results show that the frequency of *only* in CDS significantly predicts the frequency of *only* in CPS. As presented above, the overall frequency of *only* increases significantly in both CDS and CPS. Crucially, the changes in frequency of *only* pattern similarly for both child and caregiver speech. However, future research is needed to determine if frequency changes are due to caregivers being sensitive to the communicative needs of their children or vice versa. Nonetheless, understanding the dynamic aspects of both CDS and CPS is crucial when doing laboratory research and when building accurate models of child language acquisition.

## References


Stephen Crain, Weijia Ni, and Laura Conway. 1994. Learning, parsing and modularity. *Perspectives on sentence processing*, pages 443–467.

Nomi Erteschik-Shir. 1997. *The dynamics of focus structure*. Cambridge University Press.

Soyoung Kim. 2011. Focus particles at syntactic, semantic and pragmatic interfaces: The acquisition of only and even in english. *Honolulu: The University of Hawaii*.

Kung-Yee Liang and Scott L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Brian McWhinney. 2000. The CHILDES project: Tools for analyzing talk. *Mahwah, NJ*.

Kevin B. Paterson, Simon P. Liversedge, Diane White, Ruth Filik, and Kristina Jaz. 2006. Children's interpretation of ambiguous focus in sentences with "only". *Language Acquisition*, 13(3):253–284.

Dimitrios Vagenas and Vasiliki Totsika. 2018. Modelling correlated data: Multilevel models and generalized estimating equations and their use with data from research in developmental disabilities. *Research in Developmental Disabilities*, 81:1–11.