The value of corpus data in interpreting experimental results

Elena Lieven
ESRC International Centre for Language and Communicative Development
(LuCiD)
University of Manchester

Grant number ES/L008955/1
## Experiments and Corpus data

<table>
<thead>
<tr>
<th></th>
<th>Experiments</th>
<th>Naturalistic data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantages</strong></td>
<td>Can target particular constructions/forms</td>
<td>Only way to capture what children hear</td>
</tr>
<tr>
<td></td>
<td>Can pin down development</td>
<td>Possible breadth of coverage and contexts</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>Can be very artificial</td>
<td>Difficulty of sampling dense enough data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How naturalistic is it?</td>
</tr>
</tbody>
</table>

**Obvious answer:** Use one as a control on the other!
Complex sentences

– here: main clause + adverbial clause

– express a specific relationship between two or more situations

Our focus: after, before, because and if

<table>
<thead>
<tr>
<th>Situation A</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>After I put the kettle on</td>
<td>Temporal</td>
</tr>
<tr>
<td>I ate a piece of toast</td>
<td></td>
</tr>
<tr>
<td>Before she moved to Boston</td>
<td>Temporal</td>
</tr>
<tr>
<td>she lived in LA</td>
<td></td>
</tr>
<tr>
<td>Because she fell off her bike.</td>
<td>T + Causal</td>
</tr>
<tr>
<td>she grazed her knee</td>
<td></td>
</tr>
<tr>
<td>If you don’t pay the money</td>
<td>T + Conditional</td>
</tr>
<tr>
<td>I’ll turn you in.</td>
<td></td>
</tr>
</tbody>
</table>
• Complex sentences can occur in two clause orders:
  – After I put the kettle on, I ate a piece of toast. [sub-main]
  – I ate a piece of toast after I put the kettle on. [main-sub]
  – Because she fell off the bike, she grazed her knee. [sub-main]
  – She grazed her knee because she fell off the bike. [main-sub]
• The order can be **iconic** or **non-iconic**:

Clause order reflects order of events in the real world.

Clause **order is reversed** w.r.t. order of events in the real world.
Experiments on children’s comprehension

Mixed picture:

- Iconic orders are understood better than non-iconic orders
  (e.g., Blything et al., 2015; Corrigan, 1975; Emerson, 1979; Clark, 1971)

- No difference between iconic and non-iconic orders
  (e.g., Amidon & Carey, 1972; Gorrell et al., 1989)

- Information in the main clause is understood/processed
  better than information in the subordinate clause
  (e.g., Amidon & Carey, 1972; Gorrell et al., 1989; Johnson, 1975; Townsend & Ravelo, 1980)

  - Before understood earlier than after
    (Blything et al., 2015; Clark, 1971; Feagans, 1980; Goodz, 1982)

  - After understood earlier than before
    (Carni & French, 1984)

  - No difference between before and after
    (e.g., Amidon & Carey, 1972; French & Brown, 1977; Gorrell, Crain, & Fodor, 1989)

See De Ruiter et al (2018) for these references
Design — De Ruiter et al 2018, *Cognition*

- Forced-choice picture selection task
  - Instruction: *Touch the matching story after the beep!*

CHILDES.vi.2019
Results

- 4;0s barely above chance
- **Before** better than all other types
- **Iconic** orders better than non-iconic ones
- (Adults at ceiling)

De Ruiter et al. 2018
Correlations between input and language development have been found for other aspects of grammar such as morphology, but also syntax (e.g., relative clauses: Kidd et al 2007; Brandt et al. 2009).

So, what do the complex sentences that children actually hear sound or look like?
Sampling

Two important factors:

- Occurrence (frequency) of the linguistic feature
- Sample size and density of data collection:
  - Influences the probability of detecting a feature in the corpus
  - Influences the reliability of the estimates we make
  - Influences the estimated age of the first occurrence

Tomasello & Stahl, 2004
Rowland & Fletcher 2006
Types of corpora

**traditional** = 1 hour per 1-2 weeks, 26-52 hours per year = 1-2%

**high density** = 5 hours per week, 260 hours per year = 10%
**double density** = 10 hours per week = 20%

**diary** = Bowerman, Braunwald, Rowland,

Lieven & Behrens 2012
The data

• Child-directed speech (CDS) from two dense corpora of British English (Lieven, Salomo, & Tomasello, 2009)
• 6 weeks starting at the 3\textsuperscript{rd} birthday, several recordings every week:

\begin{tabular}{c|c|c|c|c|c|c}
  & W1 & W2 & W3 & W4 & W5 & W6 \\
\hline
Thomas & & & & & & \\
\hline
Gina & & & & & & \\
\end{tabular}

~ 96 hours

De Ruiter et al. in prep
Vastly different overall frequencies for different adverbials

- **after** and **before** much less frequent than **because** and **if**: both used more in other constructions but **after** almost exclusively

Relative frequencies similar for mothers and children

- **because** children >> adults (73.1% vs. 58.8%, p < .0001)
- **if** children << adults (24.4% vs. 35.2%, p < .0001)
- **before** children << adults (1.9% vs. 4.5%, p = .0003)

Children used adverbial sentences in isolated utterances more than adults - as replies in 20.4% of the cases, in contrast with only 8.4% for the mothers (p< .0001)

So even though children hear **before** much less than **because** and **if**, they are better with it in the experiment
Clause order: adults and children

Relative frequencies very similar
Iconicity was the most important factor in the experiment. But not (entirely) predicted by input frequency.

Clause order by type in CDS corpus:

Because she fell off the bike, she grazed her knee. [sub-main] iconic
➢ She grazed her knee because she fell off the bike. [main-sub] non-iconic

So even though children hear the “[main clause…], because [sub clause…]” order much more often, they find the reverse, iconic order easier to understand in the experiment.
Other findings....
Subjects

- Nominal form mostly pronouns:
  - 92.4% of main clauses
  - 86.3% of subordinate clauses

- Definite/indefinite NPs or names relatively rare but these tend to be used in experiments (e.g. *the girl, the dog*)
Patterns very similar for mothers and children

- **because** mainly used in speech act utterances
- **if** mainly used in content utterances

In experiments **if** and **because** are used in content utterances
• About 50% of all complex sentences consisted of *more than* one main clause and one subordinated clause.
Additional Complexity (2)

• Reasons for additional complexity (ex.):
  – Complement clauses
    • *What number would we dial if we wanted to talk* to Daddy at work?
  – Compound complex sentences
    • *And she wanted Smudge but Smudge wouldn't go into her house because they used to have a dog.*
  – Multiple subordinators
    • *If you drop it on the floor there'll be trouble, Thomas, because you've been warned.*
    • *They make you very, very poorly if you take these when you don't need them.*
  – Constituent coordination
    • *If you're happy and you know* it stamp your feet.
Additional complexity (3)

... And it can get very complex indeed:

- We haven't got a bird table so if I just throw the bread out on the grass and it goes soggy you eat it and that's not what you're supposed to do because when I throw the bread out it's stale.

- Well either the train's very early or Mummy's very late, because if the train was going past Mummy's house before Mummy had left her house to go to the platform then either Mummy's running very, very late and she would've missed the train anyway, or the train's going past too early.
Summary of corpus findings

• Great variation in overall frequency of prepositions/connectives, and relative frequency of use as subordinator:
  – *before*/*after* relatively rare, and mainly used in other constructions

• Type-specific preferences in clause order:
  – iconic clause order for *before*, *after* and *if*
  – non-iconic order (main-sub) for *because*

• Majority of clauses have pronominal subjects

• *because* mainly used for speech act sentences
  *if* mainly for simple content sentences, though mothers use significantly more hypotheticals

• Additional syntactic complexity common
<table>
<thead>
<tr>
<th>Results from the experiment</th>
<th>Possible explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>before</strong> better than <strong>after</strong></td>
<td><strong>after</strong> is used in more different constructions with other meanings?</td>
</tr>
</tbody>
</table>
| **before** and **after** understood better than **because/if**, despite much lower frequency of use by both adults and children | -the cognitive challenge of causality and conditionality?  
-**because** used much more frequently in speech acts  
-**if** used in hypotheticals significantly more by adults than children |
| **because** understood better in iconic (sub-main) order despite being used in main-sub order by adults | Iconicity makes for easier processing at early stages?  
Children use **because** in isolated sentences in answer to adult questions? |
| 4;0s at chance despite using these terms | No context  
-significantly above chance if context provided (de Ruiter et al., submitted) |
Effects of context

Mean accuracy in De Ruiter et al.'s (2018) study with isolated sentences (NoContext) and the current study (Context) for the four-year-old and the five-year-old group. The red dashed line indicates chance level. The error bars indicate standard error.
Conclusions

• Frequency mapping ubiquitous:
  – Can explain why children can do things early and some things late
    • Possibility of schemas – “He ,, because ,,,”

• When frequency can’t account for the results:
  – frequent in the corpus and children learn it late
    • Cognitive complexity; Pragmatic usefulness
  – infrequent in the corpus and children learn it early
    • One-to-one form-function mapping; Pragmatic usefulness; Support from other constructions

[Good] corpus data is essential to understanding development
- It can clarify and explain experimental results
- It can generate new hypotheses for testing
The team ...

Laura de Ruiter  
*Tufts University*

Anna Theakston  
*U of Manchester*

Silke Brandt  
*Lancaster University*

Our funder

Kimberley Bell  
*U of Manchester*

Heather Lemen  
*U of Manchester*

LuCiD has received funding for another 5 years under the leadership of:

Julian Pine  
*(Liverpool)*

Anna Theakston  
*(Manchester)*

Gert Westerman  
*(Lancaster)*


De Ruiter et al. (in prep) Structural and interactional aspects of adverbial sentences in English mother-child interactions: an analysis of two dense corpora


The end

Thank you!