Topics in Linguistics

Week 3. Experimental design and the responsible use of numbers (I)

Why we do experiments
- In this context, we’re generally interested in the state and developmental course of children’s linguistic knowledge.
- What does the child know?
- To what extent does it differ from what an adult knows?
- We have logical, abstract reasons to believe that a lot of what kids ultimately know about language is not deduced from their language input—but what evidence is there?

Universal Grammar
- The poverty of the stimulus
  - 1 2 3 — what’s next?
  - 1 2 3 5 — what’s next?
- Properties of “innateness”?
  - Independence from external evidence.
  - Universality?
  - Early emergence?
- Constraints and the absence of negative evidence.
  - Candidate A: Who does Arnold wanna make breakfast for? Who does Arnold wanna make breakfast?
  - Candidate B: Who does Arnold wanna make breakfast for? *Who does Arnold wanna make breakfast?

Hypotheses
- Where you have two hypotheses that make different predictions, you use an experiment to determine which predictions are actually borne out.
- A standard setup in child language studies is pitting an experimental hypothesis ($H_1$) such as “Children (Like Those We Are Testing) Know Grammatical Constraint X” against the null hypothesis ($H_0$) They Don’t.

This should be difficult
- Experiments are naturally subject to error. We’re measuring things in the real world.
- We only want to reject the null hypothesis if we’re sure.
- So, we want to “stack the deck against” $H_0$. Exclude the possibility that kids give the correct answers for the wrong reasons.

<table>
<thead>
<tr>
<th>Reject $H_0$</th>
<th>$H_0$ actually true</th>
<th>$H_0$ actually false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I error</td>
<td>Correct</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Do not reject $H_0$</th>
<th>$H_0$ actually true</th>
<th>$H_0$ actually false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>Type II error</td>
<td></td>
</tr>
</tbody>
</table>

Production and comprehension
- Two things to look at to assess kids’ grammatical knowledge.
  - Naturalistic production (e.g., CHILDES transcripts) is good for some things, but not so good for others.
  - If we’re interested in a particular construction, often this needs to be elicited in an experimental setting in order to get enough examples.
- Non-production: Constraint? Or preference?
- Test comprehension.
  - Act-out
  - Grammaticality judgment
  - Truth value judgment
Elicited production

- Wanna contraction
  - Who do you want to kiss?
  - Who do you wanna kiss?
  - I want to kiss Bill
  - I wanna kiss Bill
  - Who do you want to kiss Bill?
  - Who do you wanna kiss Bill?

  - All signs point to the last one being good. But it isn’t. Do kids know this?

  - So: try to get kids to say Who do you wanna kiss Bill?
  - Great. Suppose they don’t. Have we shown that they know the constraint?

  - *Hint: No.*

But I don’t wanna contract.

- The fact that a kid never says Who do you wanna kiss Bill? doesn’t tell us anything unless the kid would have otherwise contracted.

- So we need controls in order to determine whether the kid knows/uses wanna contraction to begin with.

  - *So: Try to get the kid to say Who do you wanna kiss? as well as trying to get them to say Who do you wanna kiss Bill?*

  - If the kid never says wanna, we have no evidence of anything.

  - Also: Maybe the kid knows all there is to know about wanna contraction, but prefers not to contract.

Here’s how it might be done.

- Design the experiment to be some kind of game, to keep the kid interested and willing to continue.

- Use a puppet. Kids are more willing to interact with the puppet, more willing to disagree with a puppet.

  - Though: Gordon (1996) relates a tale of testing Kadiveu kids, who “had never encountered puppets before and reacted with a mixture of curiosity and fear that often led to tears”

  - Ratty missed snacktime and is probably hungry. Various toy food items are arrayed in the playspace.

  - E: The rat looks kind of hungry. I bet he wants to eat something. Ask him what.

  - C: What do you want?

  - R: Huh?

  - C: What do you wanna eat?

  - R: Is that pepperoni pizza over there? I’ll have some of that.

  - E: I bet the rat wants someone to brush his teeth for him. Ask him who.

  - C: Who do you want to brush your teeth?

Analyzing the results

- Find the kids who

  - Produced both subject and object questions

  - Produced wanna sometimes.

  - H₁: Kids know that want to cannot contract to wanna over a (subject) wh-trace.

  - H₀: They don’t.

- Expectation given H₀ would be that kids would not distinguish subjects and objects; they would be as likely to contract in one case as in the other.

- If kids contract often with objects and never with subjects, that points to H₁. We could reject H₀.

Testing interpretation

- Do kids assign the same meanings to a sentence as adults? (More meanings? Fewer meanings?)

  - Constraints on meaning (e.g., Binding Theory)

- The Truth Value Judgment task is a popular way to approach this.

  - Pros:

    - Fun for the kids.

    - Minimal extra cognitive demands

    - Gets at alternative meanings an act-out task can’t reliably exclude

  - Cons:

    - A trial takes a long time, not many data points collected.

TVJ

- The idea: Set up a context by telling a story. Provide a test sentence which is either true or false of the situation (and have the puppet say it). Kid then either agrees with the puppet and rewards it, or disagrees and punishes it.

- If the puppet is wrong, the kid is asked “What really happened?”

- Also: Kids often like the puppet to be right, and will more readily agree with the puppet.

- So: stack the deck against H₀, and have adult-impossible readings correspond to “yes” responses.
Principle C, for example.

- Jumping competition (Crain & Thornton 1998)
  - This is a story about a jumping competition. The judge is Robocop. Last year he won the jumping competition, so this year he gets to be judge. This year, these guys, Cookie Monster, the Troll, and Grover are in the jumping competition. They have to try and jump over this log, the barrels, and the benches over here.
  - R: The winner of the competition gets a great prize: colored pasta! See, it’s in this barrel right here.

Against all odds

- T: No, Robocop, you’re wrong! I am the best jumper. I think I should get the prize. I’m going to take some colored pasta for myself.
- K: Let me try to say what happened. That was a story about Robocop, who was the judge, and Cookie Monster, and Grover, and there was the Troll. I know one thing that happened. He said that the Troll is the best jumper.
- C: No!! Bad Kermit. Eat this rag.
  - Yet he said that Troll, is the best jumper is true.

TVJ

- Distinguishing meaning_1 (disallowed by adult constraint) and meaning_2 (allowed by adult constraint).
- Test sentence should be true on meaning_1, false on meaning_2.
- Child judges puppet’s report to be true (reward) or false (punishment).

Experimental design

- What we’re trying to determine is the degree to which variables in the situation affect one another.
  - Does a Principle C configuration preclude a certain interpretation?
  - Does a subject unfit-extraction preclude wanna contraction?
- Independent variables are the presumed causal variables.
- Dependent variables are the presumed caused variables.
- Nuisance or confounding variables are other factors that may introduce systematic noise

Between- and within-subjects

- Between-subjects designs vary independent variables with the subjects, so each subject represents one of the values (levels) of the independent variable.
  - Age, for example ("Cross-sectional").
- Within-subjects designs vary independent variables for each subject, so each subject sees all of the levels of the independent variable.
  - Subject extraction and object extraction, for example.
  - Age, for another ("Longitudinal").
Task considerations

- Ideally, we want test items to be distinguished by just the factor we're looking at.
- This is important because other things may play a role and may confound the result.
- If we find that kids are slower on:
  - Who did Pat say met Chris?
  - Than on:
  - Who met Chris?
- Can we conclude that it takes more time to process a longer-distance extraction? Yes.
- Well, it could just take longer because there are more words—we need to rule that out if we want to conclude that it has to do with long-distance extraction.

Things that matter

- Items should include controls.
  - Ensure that the subjects are performing the task.
  - Rule out confounding variables.
  - Irrelevant items to mask the actual goal of the experiment
- Items should be presented in different orders
- ruling out another confounding variable

Things that matter

- The instructions given matter a lot. Is the task clear?
  - Circle 1 for grammatical, 5 for ungrammatical.
  - Circle 1 if the sentence sounds ok, 5 if you would never use it.
  - Give the puppet a cookie if his sentence makes sense, and give him a rag if his sentence is silly.
- Practice with feedback
  - To confirm that the subjects understand the task (and are comfortable that they do), run a couple of practice trials.
  - Practice items should not be test items. Should be relatively easy.

Things that matter

- Balance the responses
  - If a "no box" will be considered to have scored perfectly, there is a huge uncontrolled confound.
  - If you are testing for obliviousness to a constraint, but obliviousness would yield all "yes" responses (or a big preponderance), subjects may start to "second guess" themselves.
  - Fillers/controls are for this.
- Balance the items
  - There should be the same number of items at each level of your independent variable(s).
  - This maximizes the power of statistical analysis later.
  - If a subject misses one, not a huge problem, but design it as a nice "square" if you can.

Complexity

- The sandwich arrived.
- The sandwich the judge ordered arrived.
- The sandwich the judge the president appointed ordered arrived.
- The president appointed the judge who ordered the sandwich that arrived.

Complexity

- The nanny was adored by all the children.
- The nanny who the agency sent was adored by all the children.
- The nanny who the agency that the neighbors recommended sent was adored by all the children.
- The neighbors recommended the agency that sent the nanny who was adored by all the children.
What makes those first sentences so difficult?

- Some kind of processing difficulty.
- Obvious candidate (Chomsky & Miller 1963, Kimball 1973): You can’t keep track of more than two sentences at a time.
  - [The sandwich [the judge [the president appointed] ordered] arrived].
  - If at any point you need more than two verbs to finish, it’s hard.

Processing load

- The idea behind this is that the human sentence processing mechanism has some limited amount of storage capacity. It’s memory-related, in some sense.
  - (Cf. the 7 ± 2 digit span—short term memory has limits, the parser is sensitive to those/similar limits)

That’s easy enough

- The celebrity [that attacked the photographer] apologized on national TV.
- The celebrity [that the photographer attacked] applied for a restraining order.
- The first one is slightly easier, but we have no explanation for it under the “two sentences” view.
- What’s different?

Perhaps it’s “floating” θ-roles

- The celebrity that _ attacked the photographer apologized.
  - Never more than one floating θ-role.
- The celebrity that the photographer attacked _ applied…
  - At one point, two floating θ-roles.
- There seems to be something about hanging onto these nouns without having something to hook them onto. (Also sounds digit-span-like… There’s a reason phone numbers are divided).

Complexity

- The nanny [who the agency [that John recommended] sent] was adored by all the children.
  - (Thanks!) The nanny [who the agency [that you recommended] sent] was adored by all the children.
- Well, that’s funny—now what’s different?

Reference

- The nanny [who the agency [that you recommended] sent] was adored…
- The nanny [who the agency [that John recommended] sent] was adored…
- The nanny [who the agency [that the neighbor recommended] sent] was adored
- The nanny [who the agency [that they recommended] sent] was adored…
It seems like there’s a real difference—*is* there?

- Here is where the psycholinguistic experiment comes in.
- Suppose we want to test—what’s the real difference in processing difficulty between these:
  - pronouns with a referent (*you*)
  - proper names (*John*)
  - definite descriptions (*the student*)
  - pronouns without a referent (*they*)

---

Designing an experiment

- A couple of ways to go about this...
- Questionnaire:
  - The rat the cat the dog chased caught died.
  - (bad) 1 2 3 4 5 (good)
- On-line reaction time processing:
Confounds

- Who did John say that left?
- Which capybara did Madonna meet on Mars?

The point is: If you find that one sentence is judged worse than the other, we’ve learned nothing. We have no idea to what extent the that-trace violation played a role in the difference.

You want to do everything you can to be testing exactly what you mean to be testing for.

We can’t control frequency, familiarity, plausibility very reliably—but we can control for them to some extent.

Who did John say that left?
Who did John say left?

Keep everything the same and at least they don’t differ in structure, frequency, plausibility—only in that-trace. (Well, and here, length).

However—note that length now works against that-trace, unless shorter sentences are harder.

Conditions

To start, we might say we want to test two conditions:
- Sentences with a that-trace violation
- Sentences with no that-trace violation

But we can’t build these without a length confound—holding everything else constant, we still have one fewer words in the that-trace case. How do we solve this?
How can we show that the effect of the extra word that isn’t responsible for the overall effect?

The trick we’ll use is to have a second set of conditions, testing only the exact length issue. There’s no that-trace problem in object questions, so we can compare:
- Who did John say Mary met?
- Who did John say that Mary met?

To see how the difference compares to:
- Who did John say met Mary?
- Who did John say that met Mary?

Factors

We now have two “factors”—our sentences differ in terms of:
- subject vs. object question
- presence vs. absence of that

When we analyze the result, we can determine the extent of the influence of the second factor by looking at the object condition and comparing it to the (disproportionately larger) effect of the presence of that in the subject condition.

2x2 factorial design

<table>
<thead>
<tr>
<th></th>
<th>without that</th>
<th>with that</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject extraction</td>
<td>Who do you think likes John?</td>
<td>Who do you think that likes John?</td>
</tr>
<tr>
<td>Object extraction</td>
<td>Who do you think John likes?</td>
<td>Who do you think that John likes?</td>
</tr>
</tbody>
</table>

Often this is drawn in a table, with each factor on a different dimension.
This is known as a 2x2 factorial design.
Context

- It turns out that the context also seems to have an effect on people's ratings of sentences.
- What comes before can color your subjects' opinions. This too needs to be controlled for.
- One aspect of this is that we generally avoid showing a single subject two versions of the same sentence (more relevant when they're more unique than the John and Mary sentences)—the reaction to the second viewing may be based a lot on the first one.
- Another is that you want to give the sentences in a different order to different subjects.

Strategy

- You also don't want your subjects to "catch on" to what you're testing for—they will often see that they're getting a lot of sentences with a particular structure and start responding to them based on their own theory of whether the sentence should be good or not, no longer performing the task.
- Nor do you want to include people who seem to simply have a crazy grammar (or more likely just aren't understanding or doing the task).

Fillers

- The solution to both problems is traditionally to use "fillers", sentences which are not really part of the experiment.
- These can provide a baseline to show that a given subject is behaving "normally" and can serve to obscure the real "test items."
- There's no answer to "how many fillers should there be?" but it shouldn't be fewer than the test items, and probably a 2:1 (filler:test item) ration is a good idea.
- Fillers can't be all good! About half should be bad.

Instructions and practice

- Another vital aspect of this procedure is to be sure that the subjects understand the task that they are supposed to be performing (and all in the same way).
- The wordings of the instructions and the rating scales are very important, and it's a good idea to give subjects a few "practice" items before the test begins (clear cases for which the answers are provided).

Instructions

- "Is the sentence grammatical?" is not a good instruction.
  - The closest the naive subject can come to "grammatical" will probably be to evaluate based on prescriptive rules learned in grammar classes—the term does not have the same meaning in common usage.
- "Is this a good sentence?" also has problems.
  - I'd never say that, I'd say it another way.
  - That could never happen.

Numerical/category ratings

- How do you ask people to judge?
  - Good/bad
    - Forces a choice, for anything other than "certainly good" and "certainly bad" there's a chance that it doesn't reflect the subject's actual opinion—no differentiation between "great!" and "well, kind of ok"
  - Good/neutral/bad
    - Neutral also tends to get used for "I can't decide" which is different from "I'm confident it has an in-between status" (doesn't change much if you call it "in-between")
Numerical/category ratings

- Rate the sentence: (good) 1 2 3 4 5 (bad)
  - Some people will never use the ends of the scale, likely to confound certainty with acceptability. Also, for certain applications, “3” is unusable.
- Rate the sentence: (good) 1 2 3 4 (bad)
  - Can be treated as a categorial judgment, may be able to factor out some personality aspects. This is the one I tend to like best.

Online tasks

- The nice thing about an online experiment is it to some extent takes it “out of their hands.” The subject simply reacts, and we time it.
- Nevertheless, it is still important to ensure that the subject is performing the task, paying attention.
  - Often can be addressed by questions about the sentence afterwards they must answer.
  - Feedback can strengthen the motivation.

Your task (Lab #3)

- Negative questions in child English often don’t have the adult shape.
- Suppose that we’re testing the hypothesis that children have difficulty with negation on an inverted I (negation has to stay within the IP).
- Predicts: trouble with What didn’t you buy? but not necessarily with Who didn’t you meet?
  - Where he couldn’t eat the raisin?
  - What did he didn’t wanna bring to school?
  - Why can you not eat chocolate?
- Design an experiment to test the prediction.
  - What type of experiment?
  - What factors?
  - Give an example of at least two different trials.
  - What would you expect to see if the hypothesis is true?