Research Methods in Linguistics

1302740

Lecture (2)

Judgment Data
**Broad sense:** a program of applying formal experimental methodology(ies?) to investigate hypotheses relevant to current syntactic theory.

**Narrow sense:** focus on *acceptability* judgements
Acceptability judgments involve explicitly asking speakers to “judge” (i.e., report their spontaneous reaction concerning) whether a particular string of words is a possible utterance of their language, with an intended interpretation either implied or explicitly stated.
Why do we need acceptability judgments?

What are the problems with acceptability judgments?
There is no direct way to access I-language (the speaker’s knowledge of their language).

Corpus data typically
- aggregate across speakers
- include performance errors
- allow no straightforward distinction between non-occurring and ungrammatical
- may not exist
WHY DO WE NEED JUDGMENT DATA?

- provide evidence (under certain assumptions) about the grammaticality of utterances that have never been naturally produced.
- provide evidence about the status of phenomena that occur so rarely in spontaneous language use that we could not otherwise learn about them.
A further advantage of judgment data over spontaneous usage data is that the latter will include some proportion of production errors (slips of the tongue/pen/keyboard, etc.), the vast majority of which will be judged as ill-formed by the very speakers who produced them, and which therefore should not be generated by the grammar.

Unlike analyzing corpora, collecting judgments allows the researcher to question speakers about what they have said.
Intuitions about grammatical/acceptable and ungrammatical (or unacceptable) sentences indirectly reveal rules (principles) of grammars.

English: (i) *met Mary John

(ii) *loves you.
One common method in the study of perception is to ask participants to report their perceptions along some sort of scale (e.g., Stevens 1956, 1957).
data do not necessarily bear on how the human language faculty is actually constructed unless their “psychological reality” has been tested via some experimental procedure using another dependent measure, such as time, error rate, electrophysiological response,
judgements data demand awareness of language as an object of attention and evaluation – that is, metalinguistic awareness.
Formal research is controlled, objective, and systematic gathering of data.

The researcher carefully defines the things under study and what will and won't be studied.

It is systematic in that we carefully follow prescribed rules in gathering and assessing data.

It is controlled in that we carefully define, gather, and evaluate the data according to prescribed rules that can be reviewed for error.
The formal method provides a description based upon agreed upon units that can be measured and assessed for reliability, whereas the informal method describes the data based on the intuition of the researcher.
To collect linguistic data, a researcher may be conducting formal and informal open-ended interviewing on a variety of topics.

The aim of informal methods is to learn about the actual processes of selecting a particular linguistic form and abandoning other ones.

Informal data collection methods rarely follow cultural and institutional rules.

Information on these standards are mostly obtained through *participant-observation.*
five major respects in which typical informal linguistic judgment data gathering tends to differ from standard practices:

(i) relatively few speakers (fewer than ten),

(ii) linguists themselves as the participants,

(iii) relatively impoverished response options (such as just “acceptable,” “unacceptable,” and perhaps “marginal”),

(iv) relatively few tokens of the structures of interest,

(v) relatively unsystematic data analysis.
JUDGMENT TASKS

- Non-numerical (qualitative)
- Numerical (quantitative); to measure *the size of the difference*

**Tasks:**
1. Forced-Choice
2. Yes-No Task
3. Likert Scale Task
4. Magnitude Estimation task
5. Thermometer Task
participants are presented with two (or more) sentences, and instructed to choose the sentence that is most (or least) acceptable (perhaps by filling in a corresponding circle or radio button),

Figure 3.1. *An example of a two-alternative forced-choice task*
FORCED-CHOICE

**BENEFITS**
- are relatively easy to deploy
- increased statistical power to detect differences between conditions

**LIMITATIONS**
- can only indirectly provide information about the size of the difference between conditions
- provides no information about where a given sentence stands on the overall scale of acceptability
similar to the FC task in that it is primarily a qualitative task

The Y/N task is designed to answer the question, *Does this sentence belong to the yes-category or the no-category?*

In this way, the Y/N task probes the relationship between a single sentence and the two categories presented to the participant.

What do you wonder whether John bought? ○ Yes ○ No

Figure 3.2. *An example of the yes-no task*
**YES-NO TASK**

**BENEFITS**
- is relatively easy to deploy
- can be used to compare the relative difference between conditions

**LIMITATIONS**
- less sensitive than the FC task at detecting qualitative (numerical estimates) differences between two conditions
acceptability judgements are \textit{gradient}

(i) Who does John think Mary will fire?

(ii) ?*Who did Mary wonder whether they will fire?

(iii) *Who did John meet the girl who will marry?

\textit{What is the status of sentences of intermediate acceptability like (ii)?}
“an adequate linguistic theory will have to recognise degrees of grammaticalness” (Chomsky 1975).

Are such discrepancies an artefact of the absence of an unambiguous notational system and the lack of a systematic way of quantifying linguistic intuitions, or do they represent real disagreements about the acceptability of the structures in question?
A psychometric response scale and often called a ‘rating scale’,

Named after Dr. Rensis Likert, a sociologist at the University of Michigan, who developed the technique.

his goal was to develop a means of measuring *psychological attitudes* in a “scientific” way

used in questionnaires such as *attitude surveys*. 
primarily used in questionnaires to obtain participant’s preferences or degree of agreement with a statement or set of statements.

Likert scales are a non comparative scaling technique.

Respondents are asked to indicate their level of agreement with a given statement by way of an ordinal scale.
LIKERT SCALE TASK

- used for measuring the social attitude.
- Questionnaires is prepared.
- We will ask the individual to express his response.
- Some relevant and indirect statement will be used to reveal the attitude.
- Scale also specifies the crucial shades of opinion.
The likert uses several degrees of agreement or disagreement. For example:

1. *Strongly Approve*
2. *Approve*
3. *Undecided*
4. *Disapprove*
5. *Strongly Disapprove.*
What do you wonder whether John bought?

<table>
<thead>
<tr>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Each point of the scale above carries a score.

- “Strongly approve” is given the highest score (7)
- “strongly disapprove” is given the least score (1).
the method of construction of likert scale

- The researcher gathers a large number of statements which clearly indicate favourable or unfavourable judgments towards the sentence in question.
- The scores are consistently arranged either from the highest to the lowest, or from the lowest to the highest.
- The questionnaires consisting of the five points with respect to the grammaticality of a sentence are administered to the respondents who indicate their responses.
- The responses will imply various scores of grammaticality/acceptability judgment.
LIKERT SCALE TASK

- **the method of construction of likert scale**
  
  - By adding up the different scores of an individual, *the total score is calculated.*

<table>
<thead>
<tr>
<th>Answer Choices</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely satisfied</td>
<td>14.29%</td>
</tr>
<tr>
<td>Quite satisfied</td>
<td>9.52%</td>
</tr>
<tr>
<td>Somewhat satisfied</td>
<td>21.43%</td>
</tr>
<tr>
<td>Neither satisfied nor dissatisfied</td>
<td>23.81%</td>
</tr>
<tr>
<td>Somewhat dissatisfied</td>
<td>11.90%</td>
</tr>
<tr>
<td>Quite dissatisfied</td>
<td>14.29%</td>
</tr>
<tr>
<td>Extremely dissatisfied</td>
<td>4.76%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>42</strong></td>
</tr>
</tbody>
</table>
BENEFITS

- It supplies more precise and definite response towards the grammaticality\acceptability of a sentence.
- Numerical and intuitive. The former means that LS can be used to answer questions about the size of a difference between conditions by leveraging inferential statistical tests such as ANOVA.
- The likert scale permits the revelation of several (five) degree of agreement or disagreement, but YN scale is based on only two alternative responses.

LIMITATIONS

- The judgement on the basis of total score. The total score values may be the same in many cases, but the judgment may be different towards a sentence.
- Numerical problems: suggests that the intervals between points are uniform: the interval between 1 and 2 is one unit, the interval between 2 and 3 is one unit, and so on.
BENEFITS

- Previous questions may influenced responses to any further questions that have been asked.

LIMITATIONS

- Participants may not be completely honest - which may be intentional or unintentional.
- Participants may base answers on feelings toward surveyor or subject.
- Scale requires a great deal of decision-making.
- can take a long time to analyze the data
MAGNITUDE ESTIMATION SCALE (ME)

- ME is an experimental technique used to determine quickly and easily how much of a given sensation a person is having.
- In an ME experiment subjects are presented with a standard stimulus (a *modulus*) and are asked to express the magnitude by a number.
- They are then presented with a series of stimuli that vary in intensity and are asked to assign each of the stimuli a number relative to the modulus.
ME IN GRAMMATICALITY/ACCEPTABILITY JUDGMENT

- Subjects assign a number:
  - to the modulus (a standard sentence) to reflect magnitude of pertinent characteristics (grammaticality\acceptability)
  - to each successive stimulus (sentence) to indicate (rate) apparent magnitude of grammaticality\acceptability relative to the standard sentence

<table>
<thead>
<tr>
<th>Standard:</th>
<th>Who thinks that my brother was kept tabs on by the FBI?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptability:</td>
<td>100</td>
</tr>
<tr>
<td>Item:</td>
<td>What do you wonder whether John bought?</td>
</tr>
<tr>
<td>Acceptability:</td>
<td>____</td>
</tr>
</tbody>
</table>

Figure 3.4. An example of the magnitude estimation task
Scaling in ME is not about absolute accuracy of judgments; Scaling is about the relative relationships between judgments of stimuli of different levels of grammaticality/acceptability.
ME provides measurements of subjective judgment on a numerical scale which can be plotted against the objective measure of the stimuli giving rise to the judgment.

It does not restrict the number of values which can be used.
BENEFITS
- ME yields interval scales, which allow the use of parametric statistics
- Mathematical operations can be applied to the estimates, allowing:
  - a direct indication of the speaker’s ability to discriminate between more or less acceptable sentences
  - a direct measure of the strength of speakers’ preferences

LIMITATIONS
- Participants use a small set of (typically whole) numbers repeatedly, and (many or all of) the members of that set often stand in a salient relationship to one another that does not seem to depend on the stimuli (e.g., multiples of five or ten)
- Participants could not make ratio comparisons of the acceptability of two sentences.
**BENEFITS**

- Informants are enabled to express their intuitions without any restrictions of the judgment scale.
- They are asked to provide purely comparative judgments: these are relative both to a reference item and the individual subject’s own previous judgments.
- At no point is an absolute criterion of grammaticality applied.

**LIMITATIONS**

- Participants may be treating the ME task as a type of LS task, only with an open and infinite response scale.
BENEFITS

- The subjects themselves fix the value of the reference item relative to which subsequent judgments are made.
- The scale used by informants is open-ended and has no minimum division: subjects can always add a further highest score or produce an additional intermediate rating.
- The result is that subjects are able to produce judgments which distinguish all and only the differences they perceive.

LIMITATIONS
participants are given two reference sentences with associated acceptability values, such as 20 and 40 (analogous to freezing and boiling points).

participants can then choose values for target sentences along the real number line relative to those two points by treating it as a linear scale.

for example, a target whose acceptability is halfway between the acceptability of the two reference sentences would be rated 30.
Methodological issues that are particularly relevant to the design of judgment experiments.

- Instructions
- Materials:
  - Practice items,
  - Factorial designs,
  - Multiple lexicalizations,
  - Fillers,
- Sample size and statistical power
- Naive versus expert participants
DESIGNING JUDGMENT EXPERIMENTS

Methodological issues that are particularly relevant to the design of judgment experiments.

- Instructions
- Materials:
  - Practice items,
  - Factorial designs,
  - Multiple lexicalizations,
  - Fillers,
- Sample size and statistical power
- Naive versus expert participants
we want to convey to speakers that certain aspects of sentences are not of interest to us and should not factor into their responses (e.g. violations of prescriptive grammar rules)

to avoid that the sentence would\ wouldn’t actually be uttered in real life, and the truth or plausibility of its content.
to avoid the question of the sentence being understandable, since uncontroversially ungrammatical sentences are often perfectly comprehensible (e.g., *What did he wanted?*).

to instruct participants to imagine that the sentences were being spoken by a friend, and ask whether the sentences would make them sound like a native speaker of their language.
DESIGNING JUDGMENT EXPERIMENTS

INSTRUCTIONS

- to what extent the instructions of the experiment can influence the results.
- the experimenter has relatively little control over how participants choose to respond to the sentences presented to them.
- experimenters should focus on controlling the experiment (materials, fillers, etc.) rather than controlling the behavior of the participant.
In the LS task, it is common to provide anchor items for certain points on the scale, to help ensure that every participant uses the scale the same way.

- An anchor item is a single sentence token that the researcher assigns to a single point on the rating scale.
- It is not necessary to provide an anchor for every point on the scale.
Most hypotheses in linguistics are not about individual sentences but about types of sentences – that is, all sentences that have a particular structural property.

Whenever possible, it is desirable to create multiple lexicalizations of each condition (ideally eight or more) and distribute them evenly among the participants, in an effort to minimize the contribution of particular lexical items, facts about real-world plausibility, and so on, to the results.
These items technically will not be analyzed in the service of an experimental hypothesis, but they may be more profitably thought of as unannounced practice items.
These can serve at least three purposes:

1. reduce the density of the critical comparisons across the whole experiment, reducing the chances that participants will become aware that a particular sentence type is being tested, which could trigger conscious response strategies.

2. can be used to try to ensure that all the possible responses (yes and no, or points along a scale) are used about equally often. This helps to protect against scale bias, which occurs when one participant decides to use the response scale differently from other participants, such as only using one end of the scale (skew), or only using a limited range of responses (compression).

3. can be used to investigate a separate research question.
TUTORIAL
### Yes/No Forced-Choice Task

<table>
<thead>
<tr>
<th></th>
<th>The corresponding sentence is grammatical/acceptable</th>
<th>Statistical Significance (No. of Yes/No)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) What will you watch on Thursday?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2) What will he watch on Thursday?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>3) What will the man watch on Thursday?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>4) What you will watch on Thursday?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>5) What he will watch on Thursday?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>6) What the man will watch on Thursday</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>The corresponding sentence is .................</td>
<td>Extremely unnatural</td>
<td>somewhat unnatural</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>----------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Who do you feel that __ insulted Pat at the theatre?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Who do you feel that Pat insulted __ at the theatre?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Who do you feel __ insulted Pat at the theatre?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Who do you feel Pat insulted __ at the theatre?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Magnitude Estimation

<table>
<thead>
<tr>
<th>Reference sentence:</th>
<th>How good does the sentence sound acceptable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do you think the website will post pictures of the new car?</td>
<td>200</td>
</tr>
<tr>
<td>Do you think pictures of the new car will be on the website?</td>
<td></td>
</tr>
</tbody>
</table>
| What do you think the website will post [pictures of __]?
| What do you think [pictures of __] will be on the website | |