Grammaticality Judgments as Linguistic Evidence
Methodological Guidelines for Measuring Grammaticality

Brian Murphy

Language, Interaction and Computation Lab
Centre for Mind/Brain Sciences (CIMeC)
University of Trento

ESSLLI 2009, Bordeaux
Course Outline

- Notions of Grammaticality, and Current Practice in Linguistics
- Others Sources of Linguistic Evidence
- Scales for Measuring Grammaticality
- Methodology for Eliciting Judgements of Grammaticality
  - Objectives of Experimentation
  - Ideal and practical choices
  - Software
  - Materials
  - Statistics
- Theoretical Implications
Objectives of Experimentation 1

- **Validity**: are we measuring what we aim to measure?
- **Reliability**: are your data purely the result of chance?
- **Replicability**: do other researchers have enough information to ‘rerun’ your experiment?
- **Generalizability**: do your results apply to other sentences/speakers than the ones you
- **Otherwise, a field is likely to spend a lot of time following up bad leads and chasing its own tail**
Objectives of Experimentation II

- Validity: are we measuring what we aim to measure?
  - as we have seen, intuitions are most often the best measure we have of grammaticality
  - but we have to be aware of strategising, misinterpretation of instructions, subconscious biases [Armstrong et al., 1983]
Objectives of Experimentation III

- Reliability: are your data purely the result of chance?
  - this is what inferential statistics are for
Objectives of Experimentation IV

- Replicability: do other researchers have enough information to ‘rerun’ your experiment?
  - ask journals about online appendices; or put supplementary materials on your website
  - describe your methods very briefly
Objectives of Experimentation V

- Generalizability: do your results apply to other sentences/speakers than the ones you tested?
  - are your materials representative of the structure and language variety you’re interested in?
  - are your speakers representative of the speech community you’re interested in?
  - how are you dealing with confounds?
The Ideal Experiment I

All other things being equal (and of course they never are) you would like your experiment to have the following characteristics:

- As many informants as possible, minimum 10
- As many judgements for each sentence as possible, minimum 10, preferably 20
- No single participant should see the same sentence twice
- The informant should not know your hypothesis
- You know the informant’s age, profession, educational background, dialect, handedness, ...
- Test items should be interspersed with fillers
The Ideal Experiment II

- Materials should be authentic
- Context should be supplied to control interpretation
- Presentation order should be randomised
- Each informant should see a similar proportion of each type of sentence (counter-balancing)
- The experimenter should not be present during the experiment
- The informants should not be known to the experimenter
- ... and probably some other things I’ve forgotten about
What Matters More? I

Featherston 2009: Relax, lean back, and be a linguist

- plausibility of content of experimental materials
- sentence or phrasal length and complexity, for example ...
- ... referential abstractness (eg too many pronouns)
- unclear meaning
- low accessibility of intended interpretation
What Matters More? II

Brian’s pet list:

- Quantity of sentences and informants
- Clear, naturalistic instructions
- Training/calibration
What Matters More? III
What Matters More? IV

The X axis represents the number of responses, the Y axis the mean score up to that point.
Symbols represent the following sentence stimuli:

- squares: ... it may take a little longer
- diagonal crosses: I was surrounded by an endless sorrow
- diamonds: Did anyone order me a plain cheese
- triangles: Throw the idol me
- crosses: Stop thinking sex about
What Matters Less? I

Featherston

- social, educational and professional status of subjects
- dialect background of subjects (except ...)
- the sex, age, and handedness of subjects
- a degree of linguistics training among subjects (but ...)
- frequency in lexis (as long as extremes are avoided)
- speed of response required
- structure type being tested (statements, questions, dialogues)
- carrying out experiments over the net or on paper
- the precise methodology, as long as it asks subjects about their receptive responses to examples
Lab-based (roughly in order of ease of use):

- **E-Prime, Presentation**: aimed at psychologists, windowing-based, can build your first experiment in an hour with no programming skills, no particular support for linguistics, very accurate timing, no inbuilt stats; **expensive** (see if you can use a copy from your Psychology department)
- **Psychotoolbox for Matlab**: aimed at vision researchers, code/windowing based, need scripting skills, no particular support for linguistics, very accurate timing, statistics for free, *kind-of-free* (toolbox is free; many universities have campus licenses for Matlab)
- **PyEPL, PsychoPy**, both for Python: aimed at vision researchers, code based, need scripting skills, no particular support for linguistics, very accurate timing, statistics for free, free
Software II

- Internet-based (work in labs too):
  - Linguist-GRID: aimed at linguists, web-form based, very easy, DEFUNCT (but on source-forge still)?
  - LimeSurvey: aimed at social sciences, web-form/configuration-file based, user-access control, no timing, has to be installed on server, data/informant/session management, summary of results, free
  - WebExp for Java: aimed at linguists and psychologists, perfectly adequate timing, configuration file based, has to be installed on webserver, data/informant/session management, not-quite-platform-independent, free
Software III

- Paper/Mail-based, for your inner-Luddite: MiniJudge
- [Myers, 2007, Myers, 2009]: ‘small-scale experimental syntax’:
  - Experimental sentences only (no fillers)
  - Only as many sentence sets as are needed for statistical validity (around 5)
  - Only as many speakers as are needed for statistical validity (around 7)
  - All speakers get all sentences (no counterbalancing of sentence lists)
  - Binary YES/NO judgments
  - Maximum of two binary factors
  - Random sentence order, which is treated as a factor (helps control for some processing effects)
Materials

- Nowadays, authentic materials are easy to find
- Google:
  - gave * the *
- Yahoo API:
  - (gave * the *) OR (send * the *) OR (send * * a *) ...
- www.webcorp.org.uk:
  - g[i|a]v[ing|e|es|en] * to *
- corpus.leeds.ac.uk (Wacky, [Baroni et al., 2009]):
  - "give|gave|given" [pos="DT"] [pos="N.*"] "to" [pos="N.*"]
- lse.umiacs.umd.edu:
  - (S1 (VP (VB give) NP (PP (TO to) NP)))
Download and install R, probably the most important software package in science (www.cran.org)

Do a basic practical statistics course at your department of statistics or psychology

Try to understand the t-test, and linear correlation, and their non-parametric alternatives Mann-Whitney and rank correlation

Ignore theory of scales [Stevens, 1946]
  > normality of the sampling distribution only thing that matters
  > so test more sentences to improve use parametric statistics
Software Wishlist I
What Comes Last

- Notions of Grammaticality, and Current Practice in Linguistics
- Others Sources of Linguistic Evidence
- Scales for Measuring Grammaticality
- Methodology for Eliciting Judgements of Grammaticality
- Theoretical Implications
References


