Empirical methods for rationalist linguistics

James Myers, National Chung Cheng University

(1) Goals of this presentation:

- Review problems with the empirical basis of generative syntax
- Explain methods proposed in the current literature for resolving these problems
- Describe my attempts to simplify these methods to the bare minimum
- Report a first application to Chinese syntax (with Niina Zhang)
- Give a demo of MiniJudge 0.1, which automates "minimalist experimental syntax"

(2) Rationalist linguistics and empiricist psychology

- Rationalism: Knowledge isn't gained solely through the senses, by babies and/or by scientists. (Empiricism claims the opposite.)
- Linguists accept "simplifying" explanations, but psychologists prefer "cause and effect" explanations (Miller, 1990); linguists use a "mathematical" approach rather than a merely "experimental" approach (Freidin & Vergnaud, 2001).

(3) Yet syntax *is* experimental psycholinguistics, as Phillips and Lasnik (2003) emphasize:

> "Gathering of native-speaker judgments is a trivially simple kind of experiment, one that makes it possible to obtain large numbers of highly robust empirical results in a short period of time, from a vast array of languages. Any good linguistics study involves carefully constructed materials, appropriate control items, and robust and replicable results. It is only because the technique is so easy and requires no more than a notebook that it is not usually described as an 'experiment'. [...] Furthermore, it is a truism in linguistics, widely acknowledged and taken into account, that acceptability ratings can vary for many reasons independent of grammaticality...." (p. 61) (See also Chomsky, 1965; Penke & Rosenbach, 2004)

(4) An example of state-of-the-art experimental design in syntax: Li (1998)

- Carefully constructed materials: Li cites 27 Chinese examples in 9 pages, sometimes more than one for each point. Many fit into a factorial design:

Individual-denoting	∅	* 三個學生在學校受傷了。	[Li's (1)]
Individual-denoting	有	有三個學生在學校受傷了。	[Li's (3)]
Quantity-denoting	∅	三枝棍子夠你打他嗎？	[Li's (8)]
Quantity-denoting	有	* 有三枝棍子夠你打他嗎？	[Li's (17a)]

- Appropriate control items: As much as possible, Li uses examples that are matched for extra-grammatical variables (e.g., lexical content, discourse context).
- Large numbers of empirical results: Li tests multiple factual claims, including effects of scope, 都 and 有, pronominal coreference, and pronominal and reflexive binding.
- Taking extra-grammatical causes of acceptability variation into account: In footnote 3, Li notes that sentences like (1) improve in the proper discourse context.

(5) So what's the problem?

> Edelman and Christiansen (2003, p. 60): Grammaticality judgments "are inherently unreliable because of their unavoidable meta-cognitive overtones, because grammaticality is better described as a graded quantity, and for a host of other reasons." Challenges posed by judgments are reviewed in Schütze (1996) and Cowart (1997), e.g.:

- Meta-cognitive overtones: Judgments are affected by so many unexpected factors that it takes careful study to learn which ones must be "taken into account." For example, Nagata (1989) confirmed that judgments are affected not only by repetition (judging the same sentence twice) but also by the presence/absence of a *mirror*!
- Grammaticality as a graded quantity: It is often hard to make binary judgments, forcing the use of "?", "?*", "*?", etc. Some experimental syntacticians conclude that grammar itself generates gradient outputs (e.g., Keller, 2000).
- Other reasons: Generative linguists focus on the ideal speaker-listener (Chomsky, 1965), despite admitting that performance data can be "noisy." Reliable generalizations can be extracted from noisy behavioral data using statistics, but this requires multiple, independent observations. Syntactic judgments made on a single sentence pair by a single, biased speaker (i.e., the researcher) cannot be analyzed statistically (Woods et al. 1986, p. 1), so their reliability cannot be tested.

(6) Real-life examples of data problems in syntax

- "[I]t seems to be commonplace for students to disagree with judgments given in texts and the research literature." (Cowart, 1997, p. 5)
- Is "Why$_i$ do you think that he left t_i?" grammatical? Lasnik and Saito (1984) say yes, Aoun, Hornstein, Lightfoot, and Weinberg (1987) say no (see Schütze, 1996).
- Is "Gestern traf sie mich fast" grammatical in German? Meinunger (2001) says yes; Rapp and von Stechow (1999) say no.
- Can 什麼 have wide scope in 「你想知道誰買了什麼？」? Huang (1982) says yes, Xu (1990) and Chen and Pan (2003) say no.
- Is 「他能做每件我不能的事」 grammatical? Soh (2005) found that 6 out of 11 native speakers found it acceptable, while 5 found it "awkward."

(7) A set of open questions in Chinese syntax (Myers & Zhang, 2005)

- Is it possible to extract out of conjuncts and adjuncts? Grosu (1973) says yes to the former, under certain semantic conditions, while Huang (1982) says no to the latter.
- We thus predict (a-b) should be better than (c-d). But is this pattern really so clear?

a. Conjunct/Topic:	那份作業，淑芬說她先寫了，然後看了報紙。
b. Conjunct/Relative:	那份淑芬說她先寫了然後看了報紙的作業，就在那裡。
c. Adjunct/Topic:	那份作業，淑芬說她如果寫了，就看報紙。
d. Adjunct/Relative:	那份淑芬說她如果寫了就看報紙的作業，就在那裡。

- Moreover, Xu and Langendoen (1985) claim that topicalization doesn't involve extraction at all, predicting that (c) should be better than (d). Do you agree?
- Anyway, can we really trust your judgments, given that they may be influenced not only by theoretical biases but also, given that Chinese relatives are left-branching and thus center-embedded, by mere parsing effects?

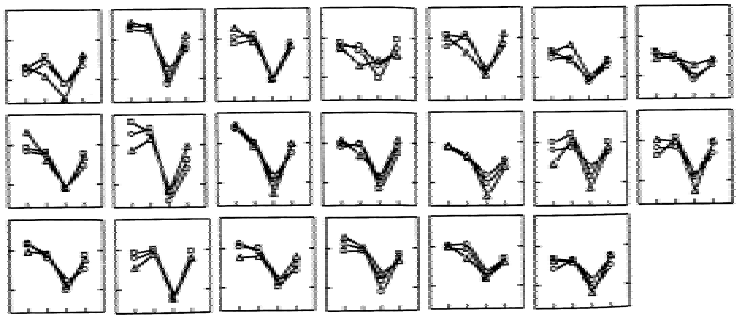
(8) Methods for dealing with such problems (Cowart, 1997)

- Meta-cognitive influences can be reduced through standard experimental procedures:
  - Randomize the order in which sentences are presented to judges
  - Mix in fillers (irrelevant sentences) so that judges won't be conscious of patterns
  - Counterbalance sentences across judges, so that no judge sees more than one item in an experimental set (thus judgments reflect individual sentences rather than sets):

Group A:	三個學生在學校受傷了。	有兩隻野狗抓住了一條蛇。
Group B:	有三個學生在學校受傷了。	兩隻野狗抓住了一條蛇。

- Gradient acceptability judgments can be measured along an ordinal scale (e.g., 1 = least acceptable, 10 = most acceptable), or via magnitude estimates as used in psychophysics (Bard, Robertson, & Sorace, 1996; Keller, 2003; Featherston, 2005b).

- The fact of linguistic variability must be dealt with even if the goal is not to study variability itself (e.g., Cedergren & Sankoff, 1974; Bod, Hay, & Jannedy, 2003) but rather to extract it out in a search for the "ideal speaker-listener."
  - Note that the gradience and variability of judgments does not necessarily reflect the nature of grammar, since these properties may arise in performance.
  - Cowart (1997) shows that highly consistent patterns emerge when variability is recognized rather than ignored (modified from his Appendix D, p. 164):



Boxes = sentence sets; lines = groups of judges; vertical axis = judgment scores (higher = more acceptable); horizontal axis = sentence type: "Who do you think {likes John / John likes / that likes John / that John likes}?"

(9) Advantages over the "Hey, Sally" method (using the term coined in Cowart, 1997, p. 2):

- Data disputes can be resolved: Featherston (2005a) established that German does indeed show *that*-trace effects, despite earlier claims to the contrary:

Wer$_i$ glaubst du, dass t_i den Schüler ausgeschimpft hat?	<	Wen$_i$ glaubst du, dass der Lehrer t_i ausgeschimpft hat?
(Who$_i$ do you think that t_i told off the pupil?)		(Whom$_i$ do you think the teacher told off t_i?)

- Unexpected facts can be revealed: Cowart (1997) found that in English, reflexives in coordinate structures may violate Principle A:

Paul requires that Cathy's parents support both himself and the child. > Paul requires that Cathy's parents support himself.

(10) Another benefit (further in the future, perhaps) is that shared methodological standards across the cognitive sciences would facilitate interdisciplinary cooperation, necessary to give substance to the notion of language as a "mental organ."

(11) Drawbacks of applying "laboratory" methods:

- Many more sentences and judges are tested than is typical in syntax
- The standard design and analysis procedures are unfamiliar to most syntacticians
- The methods are thus hardly "trivially simple": it may take weeks to test all of the factual claims in a paper like Li (1998) in the service of just one theoretical claim.

(12) Can we make it easier?

- Cowart (1997) provides an extremely clear step-by-step guide for syntacticians with no previous experience in running "laboratory" experiments.
- WEXTOR (psych-wextor.unizh.ch/wextor/en/index.php) provides general tools for guiding the design of experiments on the Web (Reips & Neuhaus, 2002).
- WebExp (www.webexp.info) provides similar tools and is the standard research tool of some prominent experimental syntacticians (Keller et al., 1998).
- ESS (Experimental Syntax Server) will be a service designed specifically for syntax experiments, including tools for sentence selection and statistics (Cowart et al. 2005).

(13) But what is the *minimum* amount of reform needed for there to be practical improvements in syntactic methodology? After all, the differences between the "Hey, Sally" and "laboratory" methods are quantitative, not qualitative:

- Syntacticians already use controls and factorial designs, albeit not systematically.
- They already test judgments across sentences and speakers, albeit usually only when they are unsure of their own judgments.
- They already know that statistics can be helpful (e.g., the 5 vs. 6 of Soh, 2005).

(14) A proposal for a minimalist experimental syntax

- Binary judgments (a forced-choice task: no "?" allowed).
- As few sentences and speakers as will permit statistically valid conclusions.
- The option to not use fillers or counterbalancing.
- Computer-assisted sentence selection.
- Computer-assisted collection of judgments.
- Fully automated statistical analyses.

(15) Making the proposal concrete

- The program should be freely available on the Web and should use statistics designed for categorical data (see, e.g., Agresti, 1996).
- A predecessor: GoldVarb (www.york.ac.uk/depts/lang/webstuff/goldvarb), a standard tool in sociolinguistics (Young & Bayley, 1996; Sankoff, 1998; Robinson, Lawrence, & Tagliamonte, 2001). Unfortunately it doesn't run experiments, is not user-friendly, requires a lot of data for statistical validity, and treats speakers as "fixed" effects.
- Thus we need a new special-purpose program for binary sentence judgments.

(16) The central question: How likely is a given judgment pattern to arise from chance, rather than being something that requires an explanation (grammatical or otherwise)?

- Researchers hope this chance probability (p) is really low; by convention, if $p < .05$, the pattern is statistically significant (though not necessarily important in practice).
- In these two simple cases, p is relatively easy to calculate, since random judgments would be like flipping a coin: heads = acceptable, tails = unacceptable.
- With one sentence pair (A & B, where A is hypothesized to be grammatical) and one speaker, there are only four possible outcomes (A is/isn't OK × B is/isn't OK).
- The probability of chance leading to a "desired" outcome (A is OK, B is bad), relative to a "reversed" outcome (A is bad, B is OK), is 50% ($p = .5 > .05$, not significant).
- Thus one judgment from one speaker can't provide enough statistical power.

(17) With a single speaker, how many sentence pairs are needed for statistical power?

- If there are n outcomes where A & B give different judgments, and a is the number of desired outcomes, the question is: What's the probability of a heads in n coin flips?
- This can be computed with a simple formula (McNemar's exact test); if the number of reversed outcomes exceed the bolded numbers below, the pattern is not significant:

Number of A > B outcomes	1	2	3	4	5	6	7	8	9	10
Max. number of A < B outcomes	*	*	*	*	*	**0**	**0**	**1**	**1**	**2**
Number of A > B outcomes	11	12	13	14	15	16	17	18	19	20
Max. number of A < B outcomes	**2**	**3**	**4**	**4**	**5**	**5**	**6**	**7**	**7**	**8**

* Pattern cannot be significant no matter how many desired outcomes there are

- Note, however, that unlike coins, speakers have memories (and perhaps theoretical biases). Thus judgments will be correlated across sentence pairs, meaning that at best, the above table underestimates the number of sentences that should be tested.

(18) With a single sentence pair, how many speakers are needed for statistical power?

- If each speaker judges both sentences, the situation is the same as in (17): the minimum number of speakers that could possibly give a significant result is six.
- Because a speaker's two judgments will influence each other (e.g., judging A as OK may bias towards judging B as bad), you might ask each speaker to judge just one sentence. Then p depends on four numbers: the total number of A judges, the number of these who judge A as OK, and similarly for B (Fisher's exact test).
- The minimum number of speakers who could give a significant result is seven, if all four (or three) A judges say "OK" and all three (or four) B judges say "bad."
- Reality is rarely so convenient; Schmitz and Schröder (2002) tested 47 speakers this way, got judgments that ran over 60% in favor of their hypothesis, and the results still didn't quite reach statistical significance ($p = .08$).

(19) What about testing several sentences and several speakers at the same time?

- Advantages: This is standard in psycholinguistics, since it deals with both sorts of noise (i.e., in Cowart's graphs in (8), both the boxes and the lines).
- Disadvantages: Categorical data make this type of analysis hard to compute, and even the soundest method (applying "GEE" to "categorical GLM models"; Agresti, 2002) is not yet used in psycholinguistics. It also requires a lot of data, but analyzing small data sets is even more computationally difficult (e.g., exact logistic regression).

(20) But what if we combine (17) with the method Lorch and Myers (1990) give for dealing with item and speaker variability (in gradient rather than categorical data)?

- For each speaker, run a logistic regression. This is the statistical heart of GoldVarb which can discover the relationship between input variables (e.g., hypothesized grammatical status) and the binary output variable (i.e., acceptable vs. not).
- The regressions then show the separate influences (some positive, some negative) of each input variable for each speaker.
- For each input variable, count the positive and negative influences across speakers.
- Finally, use the method in (17) to compute p values for each input variable.

(21) An intuitive justification of this method

- In the simplest case there is only one binary input variable, say [±G], which in a factorial experiment is distributed evenly across the items (i.e., exactly half are [+G]).
- For any given speaker, if the number of "OK" responses for [+G] is greater than the number of "OK" responses for [-G], then she shows a positive influence of [G]; if the pattern is reversed, she shows a negative influence of [G].
- If the speaker judges randomly, she will have a 50% chance of being a "positive speaker" for [G]. Thus each speaker is a "coin," and the method in (17) applies.

(22) Caveats

- I'm not yet sure if this logic still applies with multiple input variables and/or gradient input variables, though I don't see why it shouldn't.
- Logistic regression gives numerical values for the input influences, not just positive/negative signs. But these values are invalid with small data sets and also when inputs predict outputs perfectly (!). So I'm not sure how reliable the signs are.
- Counterbalancing complicates the statistics (see Raaijmakers et al., 1999, if you dare).

(23) Myers and Zhang (2005) applied this approach to the question of extraction from adjuncts in Chinese, introduced above in (7). We tested the following hypotheses:

❶ Extraction from adjuncts violates the grammar. Thus adjunct-extraction sentences (coded as [+Adjunct]) should tend to be judged worse than [-Adjunct] sentences.
❷ Sentences with topicalization are easier to parse than sentences with relativization, since only the latter creates center-embedding. Thus sentences with topicalization (coded as [+Topic]) should tend to be judged better than [-Topic] sentences, for extra-grammatical performance reasons.
❸ When judging a sentence list, the influence of grammar is fixed but parsing gets gradually easier. Thus the grammatical factor [Adjunct] should not interact with sentence [Order], whereas the parsing factor [Topic] should. (See also Snyder, 2000.)
❹ Topicalization in Chinese is movement, just like relativization. Thus there should be no interaction between [Adjunct] and [Topic], since adjuncts should always disallow extraction.

(24) Methods:

- 8 sets of 4 sentences each (including (7a-d)), creating a list of 32 sentences.
- 20 graduate students were each given the list in a different random order on paper.
- Their binary judgments analyzed assuming various combinations of input variables.
- The whole procedure, from inspiration to completion, took a day and a half.

(25) The experiment was too ambitious to be conclusive, but the results were still intriguing:

- ❶ Though [Adjunct] was not quite significant when tested by itself (p = .12), it was when [Topic] or [Order] were factored out (p = .04). Its effect was consistently negative, as predicted for a violation of grammar.
- ❷ [Topic] was significant when tested by itself or when [Adjunct] and/or [Order] were factored out (p = .04). Its effect was consistently positive, as predicted due to its relative ease of parsing.
- ❸ [Order] did not interact consistently with either [Adjunct] or [Topic]. However, the lowest p value (p = .12) was found with [Order]×[Topic], when tested alone, and this interaction was positive, implying that [+Topic] tended to improve in acceptability over the course of the list, as predicted to be the case for an parsing effect.
- ❹ No interaction between [Topic] and [Adjunct] was found (p > .26), as predicted by the hypothesis that adjunct extraction is equally bad for both topicalization and relativization. However, the high p values just mean that we failed to find an interaction, not that we succeeding in proving that there is no interaction.

(26) The most important lesson from this experiment:

- "Hey, Sally" methods would have failed utterly, given the huge amount of noise: one judge accepted almost all sentences, another rejected almost all sentences, and the rest fell in between, with none showing any obvious pattern in isolation.

(27) And now, **MiniJudge 0.1**, an Excel files with macros (巨集) that perform two key tasks:

- MiniJudge helps create sentence sets involving up to two binary input variables and puts them into randomly ordered lists for printing or emailing to judges.
- Then it carries out the procedure in (20), automatically testing for an interaction if two input variables are used (the logistic regression algorithm at its heart is taken from John Pezzullo's program at members.aol.com/johnp71/logistic.html).

(28) Take-home message

- There's no contradiction in rationalist theoreticians depending on solid empirical data.
- Psychologists should stop teasing syntacticians, since at least they run experiments, unlike phonologists!
- Syntacticians should nevertheless try to find some time in their busy schedules to improve their data-collection methodology.
- If they do so, their research will directly benefit by being empirically more solid.
- Generative syntacticians who are ready to go beyond the quick-and-dirty "Hey, Sally" method will need help from experimental psycholinguists. So just ask.

Selected references (full list at www.ccunix.ccu.edu.tw/~lngmyers/ExpSynRefs.txt)

Aoun, J., Hornstein, N., Lightfoot, D., & Weinberg, A. (1987). Two types of locality. *Linguistic Inquiry, 18* (4), 537-577.

Bod, R., Hay, J., & Jannedy, S. (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Cedergren, H. J., & Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language, 50* (2), 333-355.

Chen, L., & Pan, N. (2003). The categorical status of finite complements of xiangxin 'believe' and renwei 'think' in Chinese. In Yen-Hwei Lin (Ed.) *Proceedings of the Fifteen North American Conference on Chinese Linguistics*, pp. 45-53.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. London: Sage Publications.

Cowart, W., Hammond, M., Myers, J., Alcock, K., & Hines, J. (2005, March). *Experimental Syntax Server: Some questions.* Paper presented at 18th Annual CUNY Sentence Processing Conference, Tucson, Arizona.

Edelman, S., & Christiansen, M. H. (2003). How seriously should we take Minimalist syntax? *Trends in Cognitive Science, 7* (2), 60-61.

Featherston, S. (2005a). *That*-trace in German. *Lingua, 115* (9), 1277-1302.

Featherston, S. (2005b). Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua, 115* (11),1525-1550.

Freidin, R., & Vergnaud, J.-R. (2001). Exquisite connections: Some remarks on the evolution of linguistic theory. *Lingua, 111*, 639-666.

Grosu, A. (1973). On the nonunitary nature of the coordinate structure constraint. *Linguistic Inquiry, 4*, 88-92.

Huang, J. 1982. *Logical relations in Chinese and the theory of grammar*. MIT Ph.D. diss.

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD dissertation, University of Edinburgh.

Keller, F. (2003). A psychophysical law for linguistic judgments. In R. Alterman & D. Kirsh (eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pp. 652-657. Boston.

Keller F., Corley M., Corley S., Konieczny L., & Todirascu A. (1998). WebExp: A Java toolbox for Web-based psychological experiments. Human Communication Research Centre, University of Edinburgh. www.hcrc.ed.ac.uk/web_exp/doc/users_guide.html

Lasnik, H., & Saito, M. (1984). On the nature of proper government. *Linguistic Inquiry, 15* (2):235-289.

Li, Y.-H. A. (1998). Argument determiner phrases and number phrases. *Linguistic Inquiry, 29* (4), 693-702.

Meinunger, A. (2001). Restrictions on verb raising. *Linguistic Inquiry, 32*, 732-740.

Miller, G. A. (1990). Linguists, psychologists and the cognitive sciences. *Language, 66* (2), 317-322.

Myers, J., & Zhang, N. (2005). *Extraction from adjuncts in Chinese: An experiment in minimalist experimental syntax*. National Chung Cheng University ms. Accepted at 18th Annual CUNY Sentence Processing Conference, Tucson, Arizona, but withdrawn.

Nagata, H. (1989). Effect of repetition on grammaticality judgments under objective and subjective self-awareness conditions. *Journal of Psycholinguistic Research, 18* (3), 255-269.

Penke, M., & Rosenbach, A. (2004). What counts as evidence in linguistics? An introduction. *Studies in Language, 28* (3), 480-526.

Phillips, C., & Lasnik, H. (2003). Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Science, 7* (2), 61-62.

Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, and Computers, 34*, 234-240.

Schmitz, H.-C., & Schröder, B. (2002). On focus and VP-deletion. *Snippets, 5*, 16-17.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry, 31*, 575-582.

Soh, H. L. (2005). Wh-in-situ in Mandarin Chinese. *Linguistic Inquiry, 36* (1), 143-155.

Xu, L. (1990). Remarks on LF Movement in Chinese questions. *Linguistics, 28*, 355-382.

Xu, L., & Langendoen, T. (1985). Topic structure in Chinese. *Language, 61*, 1-27.