# An Experiment in Minimalist Experimental Syntax

James Myers
National Chung Cheng University
[Submitted March 5, 2006]]

Phillips and Lasnik (2003:61) are right to emphasize that "[g]athering 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." The value of a methodology is not measured by its complexity, and informally collected judgments are clearly powerful enough to discriminate among a wide variety of theoretically interesting hypotheses. At the same time, however, nobody would demand that empirical robustness should be sacrificed merely to maintain the methodological status quo. Linguists have worried for a long time about the ambiguous status of many informal judgments: judgment disagreements are a familiar occurrence in the syntax classroom (Cowart 1997), theoreticians like Chomsky (1981:290) have puzzled over the implications of judgment "haziness", and debates over judgments (whether they are factually correct, and if they are, whether they really reflect syntactic competence) are common at conferences, in peer reviews, and in the published literature (Schütze 1996).

In response, so-called experimental syntacticians advocate data-collection methods more in accord with the rest of cognitive science: multiple stimuli and subjects (naive speakers rather than the bias-prone experimenters themselves), systematic controls, factorial designs, continuous response measures, filler items, counterbalancing, and statistical analysis. Judgments collected in this more careful way have reconfirmed some widely accepted phenomena, such as *that*-trace effects in English (Cowart 1997), but they have also revealed hitherto unsuspected complexity, including German *that*-trace effects (Featherston 2005a) of such subtlety that informal methods had failed to detect them (see also Bard et al. 1996, McDaniel and Cowart 1999, Keller 2000, Mayo, Corley, and Keller 2005, and Featherston 2005b).

Yet rigorous experimentation requires time, technical expertise, and financial support unavailable to the average syntactician. Most seriously, time in the lab means less time for theory. This trade-off is clear when Cowart 1997 is compared with Chomsky and Lasnik 1977: in the former, great effort is expended simply to establish the existence of *that*-trace effects in naive speakers, whereas in the latter, these effects are just one of many observations in a much more ambitious analysis. What seems to be needed, then, is a middle ground between the status quo and full-blown laboratory experimentation, where methods are powerful enough to yield statistically valid results, yet are simple and cheap enough to learn and apply quickly: a minimalist experimental syntax.

This squib describes a minimalist experiment on Chinese syntax. This single simple experiment provides insights into multiple issues at a level of detail far beyond what is available from informal methods. Specifically, the experiment compares adjunct islands and conjunct islands, tests an analysis of Chinese topicalization, and studies satiation effects, where judgments weaken over time, a phenomenon claimed to provide a new window into syntactic competence and performance (Snyder 2000, Hiramatsu 2000).

## 1. Problems with informal Chinese judgments

Chinese is an apt choice for this experiment in experimental methodology for at least two reasons. First, it is one of those "vast array of languages" unlikely to be known by most readers, so to make empirical claims about Chinese convincing, we must use what Cowart (1997) calls "objective" methods. Second, judgments in Chinese are notoriously controversial, perhaps more so even than in English. For example, Huang (1982:267) claims that in (1) (his (198)), *shenme* ("what") can have wide scope. This judgment has been rejected by a number of Chinese linguists and the native speakers they consulted (including speakers from the same dialect region as Huang), such as Tang (1984), Lee (1986), Xu (1990), and Chen and Pan (2003). Xu (1990, 1996) and Shi (1994) challenge other judgments in Huang 1982. Some of the Chinese judgments in Aoun and Li 2003 also seem problematic; Ou (2006) reports disagreements by native speakers from the same dialect region as Li (e.g. their (2b), p. 133). The related empirical challenge of judgment haziness is illustrated by Soh (2005), who reports that of eleven Chinese speakers consulted on a sentence, six accepted it without hesitation while five did not (p. 151, fn 9).

(1) Ni xiang-zhidao shei mai-le shenme?[1]
you wonder who buy-ASP what
"What is the x such that you wonder who bought x?"

The syntactic claims addressed in this squib also face empirical challenges, though some partly involve the proper interpretation of data, not the mere factuality of data. According to Huang (1982), extraction from sentential adjuncts is disallowed in Chinese, thus motivating the Condition on Extraction Domain (CED). One of his examples is shown in (2) (his (33), p. 466), where topicalization out of the *yinwei* "because" clause is claimed to induce unacceptability.

(2) Zhangsan$_i$, [Lisi [yinwei [wo meiyou qing $e_i$]] hen bugaoxing].
Zhangsan Lisi because I not invite very unhappy
"Zhangsan$_i$, Lisi was very unhappy because I did not invite $e_i$."

While the CED is not controversial as an empirical generalization (see e.g. Nunes and Uriagereka 2000), the status of conjunct islands is less clear. Following Grosu (1973), Lakoff (1986), Culicover and Jackendoff (1997), among others, Zhang (2006) argues that the Coordinate Structure Constraint (CSC) of Ross (1967) can be violated with semantically related sentential conjuncts, as with the coordinator *yushi* "and" in (3) (her (86b), p. 44).

(3) Zhe jiu shi [Akiu mang-le yi zheng tian yushi xie-chulai $e_i$]
this just be Akiu busy-ASP one whole day and write-out
de wenzhang$_i$.
MOD article
"This is the article$_i$ that Akiu was busy for the whole day and wrote $e_i$."

The CSC can be preserved in the face of sentences like (3) if morphemes like *yushi* are not considered true coordinators (as Zoerner (1995) claims for similar sentences in English). Nevertheless, the general air of controversy over empirical claims in the Chinese literature suggests that it would be wise to ask for reconfirmation of the judgments claimed for sentences like (2) and (3). In particular, these judgments were made by bilingual linguists immersed in a literature that ascribes the judgments of the parallel English sentences to universal principles, so it is conceivable that Huang and Zhang were unconsciously swayed by theoretical bias. Moreover, the apparent circumvention of the CSC in (3) plants a seed of doubt about the CED, small though it may be.

Before we rush to test sentences (2) and (3) on naive speakers, however, we must first address another problem: these sentences differ in many ways besides the adjunct vs. conjunct contrast. Thus any judgment differences may actually reflect confounding variables like word frequency, verb class, or irrelevant syntactic differences. One such syntactic difference is particularly worrisome: topicalization vs. relativization. Xu and Langendoen (1990) have argued that Chinese topics are base-generated, since topics may be associated with empty positions within islands. An example is shown in (4) (their (63c), p. 15). If Xu and Langendoen are correct about topicalization, the CED would be irrelevant for (2), since it would not involve extraction. By contrast, it appears that nobody has explicitly argued against a movement analysis for Chinese relative clauses (the open question concerns rather what is moved, with Aoun and Li 2003 arguing for the NP head rather than the more commonly assumed null operator).

(4) Zhe-ge wenti$_i$ [wo conglai mei yudao-guo [neng huida $e_i$ de [ren]]]
this-CL question I ever not meet-ASP can answer MOD person
"This question$_i$, I have never met a person who can answer $e_i$."

To tease apart these issues, we should test sentences varying only in the binary factors [±Adjunct] and [±Topic]. Moreover, to let judges know that the topics are sentence-initial, the target clauses should be embedded, as in (5a-d) (respectively [+Adjunct, +Topic], [+Adjunct, -Topic], [-Adjunct, +Topic], [-Adjunct, -Topic]).

(5) a. Na-fen zuoye$_i$, [Shufen shuo [ta ruguo xie-le $e_i$, jiu kan baozhi]].
that-CL homework Shufen say she if write-ASP then read newspaper
"That homework$_i$, Shufen said if she wrote $e_i$, then she'll read the newspaper."

b. Na-fen [Shufen shuo [ta ruguo xie-le $e_i$ jiu kan baozhi]
that-CL Shufen say she if write-asp then read newspaper
de zuoye$_i$], jiu zai nali.
MOD homework just is there
"That homework$_i$ [Shufen said [if she wrote $e_i$ she'll read the newspaper]] is there."

c. Na-fen zuoye$_i$, [Shufen shuo [ta xian xie-le $e_i$, ranhou kan-le baozhi]].
That-CL homework Shufen say she first write-ASP and read-ASP newspaper
"That homework$_i$, Shufen said she first wrote $e_i$, and then read the newspaper."

d. Na-fen [Shufen shuo [ta xian xie-le $e_i$ ranhou kan-le baozhi]
that-CL Shufen say she first write-ASP and read-ASP newspaper
de zuoye$_i$], jiu zai nali.
MOD homework just is there
"That homework$_i$ [Shufen said [she first wrote $e_i$ then read the newspaper]] is there."

If Huang and Zhang are both right, we predict that [+Adjunct] sentences like (5ab) should be worse than [-Adjunct] sentences like (5cd). If Xu and Langendoen are right, and topics (but not relative clauses) are base-generated, we predict an interaction between the [Adjunct] and [Topic] factors: [+Adjunct, -Topic] sentences like (5b) should be worse than [-Adjunct, -Topic] sentences like (5d), but [+Topic] sentences like (5ac) should not differ.

There is yet one more confound to worry about: parsing. Chinese relative clauses as in (5bd) create center-embedded structures, which, as is well known, are difficult to parse despite being grammatical (Chomsky 1965). Thus sentences like (5bd) may be judged worse than sentences like (5ac) even if both types involve extraction. We are justified in ascribing a [Topic] effect to parsing if we fail to find an interaction between [Adjunct] and [Topic], since by itself, the [±Topic] contrast does not represent a grammatical difference.

Could we also exploit the speculation of Snyder (2000:580) that satiation "reflects limitations on sentence processing" rather than competence? Perhaps not; informally at least, center-embedded structures seem to sound bad no matter how much practice one gets with them. Nevertheless, Snyder's more general suggestion to add satiation effects to the syntactician's arsenal of diagnostic tools remains worthy of exploration. In this regard, it is relevant that in her judgment experiments, Hiramatsu (2000) found no evidence of adjunct island satiation in English, in contrast to other island effects that did satiate.

The complexity of the sentences in (5) blurs judgments about them, as Chinese readers may confirm informally for themselves, and of course satiation is undetectable without testing multiple sentences. Trivially simple methods are simply inadequate here.

## 2. A minimalist experiment

Of all the methodological devices listed in the introduction, only two are crucial: controls and statistics. They are two sides of the same coin, since an experiment is essentially a tool for discriminating between sources of variability. That is, it looks for correlations between the output variable (judgments) and the key input variables ([Adjunct], [Topic]), even when nuisance variables are taken into account. Making this work requires matching as many of the nuisance variables as possible, or at least distributing their variation as evenly as possible so their effects can be factored out later. Some nuisance variables have been matched in the sentence set in (5), but in order to factor out other irrelevant properties of sentence sets and speakers, more than one of each must be tested. Syntacticians already do this informally when they consider multiple sentences before choosing the "clearest" examples to present, and when they double-check judgments with their colleagues (in what Cowart (1997:2) calls the "Hey, Sally" method). In this experiment these principles were merely applied more systematically.

To generate further sets of sentences, the sentences in (5) were first doubled by replacing *ruguo* "if" in the [+Adjunct] sentences with *yinwei* "because", and *xian ... ranhou* "first ... and then" in the [-Adjunct] sentences with *budan ... erqie* "not only ... (but) also". These were then quadrupled by replacing the DP *Shufen* (a name) and the VPs *xie na-fen zuoye* "write that homework" and *kan baozhi* "read the newspaper" with syntactic equivalents. The result was a mere 32 sentences, a short list by psycholinguistic standards. Then sentence order was randomized separately for each survey form. This distributed order relatively evenly across speakers, so that any order effects (e.g. fatigue or biasing induced by prior exposure to certain sentences) could be distinguished from effects of the syntactic factors. Moreover, including actual (randomized) order in the analysis made it possible to test for satiation effects by looking for interactions with order, since satiation involves a modulation of a factor's effect over time

(stronger earlier, weaker later).

Creation of survey forms was done semi-automatically in a spreadsheet program following the step-by-step guide in Cowart 1997, and took only an hour or two. Each printed survey, with all sentences listed on one side of a piece of paper, asked for the binary good/bad judgments most familiar to linguists (skipping sentences was not allowed). This made the judgments very easy to generate and collect (though even easier would have been to distribute survey forms by email rather than using hard copies, which required retyping the results for statistical analysis). The 20 Chinese native speakers were graduate students in a linguistics program in Southern Taiwan. They were familiar with the notion of acceptability judgments but knew nothing about the theoretical issues examined here. Each survey took about ten minutes to complete, either as an in-class exercise or soon afterwards. The students were repaid by teaching them about the experiment's goals, methods, and results.

The final step was to analyze the results with a statistical method conceptually akin to the AN(C)OVA (analysis of (co)variance) ubiquitous in psycholinguistics, but designed for binary rather than continuous data (see Appendix). The entire process, from initial conception of the experiment to initial statistical analysis, took a day and a half.

The most basic analysis tested just the binary factors [Adjunct] and [Topic] and their interaction, as in ANOVA. This revealed a statistically significant negative effect of [Adjunct] on judgments: as predicted by Huang and Zhang, extraction from adjuncts was indeed worse than extraction from conjuncts. There was also a significant positive effect of [Topic]: topicalized structures were judged better than relativized structures. However, contrary to what Xu and Langendoen might predict, there was no sign of any interaction between the two factors: adjunct extraction was equally bad for [+Topic] and [-Topic] structures. Since this interaction was nonsignificant, it was removed from further analyses.

Next, order was added to the analysis as a continuous variable (as in ANCOVA). Order itself had a positive effect (overall acceptability increased over the course of the experiment), but more relevant were its interactions, which test for satiation. [Topic] did not show a significant interaction with Order (though including the interaction factor made the [Topic] effect lose significance without reducing in magnitude). By contrast, there was a significant positive interaction between order and [Adjunct]: acceptability increased only for [+Adjunct] sentences. Thus the adjunct island effect satiated in Chinese, unlike what Hiramatsu (2000) found for English. The final analysis removed the nonsignificant interaction factor, so that [Topic] again became significant (positive), and the significant effects of [Adjunct] (negative), order (positive), and their interaction (positive) remained.

To find out if these effects would have been detectable with fewer speakers, the final analysis was rerun on random subsets (100 tries per subset size). With 10 speakers, the detection

rate of the [Adjunct] effect was over 50%, and by 15 speakers it was up to 90%. The detection rate for all four effects was over 50% by 19 speakers. By contrast, when the "Hey, Sally" method was simulated by analyzing the data from all 20 speakers for just the four sentences in (5), nothing was significant (cf. also Soh's (2005) null results above).

## 3. Conclusions

This experiment was designed, run, and analyzed very quickly, yet it was still powerful enough to provide justifiable verdicts on multiple claims discussed in the theoretical literature. According to this experiment, the CED captures a true generalization about Chinese adjuncts; the CSC can be violated (assuming sentential conjuncts are true conjuncts); Chinese topicalization involves movement (as do relative clauses); the CED satiates but center-embedded structures don't, so satiation is not a diagnostic of parsing as opposed to grammar. Like all experiment reports, this squib describes a specific event, so further testing, particularly by skeptics, would be most welcome. Nevertheless, this experiment in experimental methodology has hopefully demonstrated the benefits to syntax of upgrading from the trivially simple to the merely simple.

## Appendix

Statistical analysis used GLMM (generalized linear mixed modeling; see Agresti et al. 2000). The key advantage of GLMM over more familiar categorical data tests (chi-square or logistic regression) is that it can factor out cross-subject variation, like repeated-measures ANOVA (the matched materials made by-item analysis unnecessary here; see e.g. Raaijmakers et al. 1999). Like ordinary regression, GLMM estimates the influence of input variables on the output variable, with the sign of each estimate indicating the direction of the influence. The probability that effects as strong as those observed could be due to chance is represented by the *p* value; $p < .05$ conventionally indicates significance.

Judgments were saved in a tab delimited text file, with one column each for speakers (1-20), judgments (0 vs. 1), [Adjunct] (1 vs. -1), [Topic] (1 vs. -1), sentence order (1-32), and sentence number (1-32). GLMM was run in the statistics program R (see e.g. Crawley 2005, Johnson 2004), free from www.r-project.org. The materials, judgment data, R code, and raw output of the R analyses are at www.ccunix.ccu.edu.tw/~lngproc/minexp.htm.

**References**


Agresti, Alan, James G. Booth, James P. Hobert, and Brian Caffo. 2000. Random-effects modeling of categorical response data. *Sociological Methodology* 30:27-80.

Aoun, J., and Y.-H. Audrey Li. 2003. *Essays on the Representational and Derivational Nature of Grammar: The diversity of wh-constructions*. Cambridge, MA: MIT Press.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language, 72* (1), 32-68.

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 Fifteenth North American Conference on Chinese Linguistics,* pp. 45-53.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N., & Lasnik, H. (1977). Filters and control. *Linguistic Inquiry, 8,* 425-504.

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

Crawley, Micheal J. 2005. *Statistics: An introduction using R*. Wiley.

Culicover, Peter W. and Ray Jackendoff. 1997. Semantic subordination despite syntactic coordination. *Linguistic Inquiry* 28:195-217.

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.

Grosu, Alexander. 1973. On the nonunitary nature of the coordinate structure constraint. *Linguistic Inquiry* 4:88-92.

Hiramatsu, Kazuko. 2000. *Accessing linguistic competence: Evidence from children's and adults' acceptability judgements*. Doctoral dissertation, University of Connecticut, Storrs.

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

Johnson, K. (2004). *Quantitative methods in linguistics*. UC Berkeley ms. Available at http://corpus.linguistics.berkeley.edu/~kjohnson/quantitative/

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

Lakoff, George. 1986. Frame semantic control of the coordinate structure constraint. Anne M. Farley et al., eds., *Chicago Linguistic Society 22, Part 2: Papers from the Parasession on Pragmatics and Grammatical Theory,* pp. 152-167. Chicago: CLS.

Lee, H. Thomas. 1986. *Studies on quantification in Chinese*. PhD dissertation, UCLA.

McDaniel, Dana and Wayne Cowart. 1999. Experimental evidence for a minimalist account of

English resumptive pronouns. *Cognition* 70: B15-B24.

Mayo, N., Corley, M., & Keller, F. (2005). WebExp2 experimenter's manual. Available online at www.webexp.info.

Nunes, J., and J. Uriagereka. 2000. Cyclicity and extraction domains. *Syntax* 3 (1): 20-43.

Ou, Tzushan. 2006. *Suo relative clauses in Mandarin Chinese*. National Chung Cheng University MA thesis.

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

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*, 416-426.

Ross, John Robert. 1967. *Constraints on variables in syntax.* Ph.D. dissertation, MIT.

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

Shi, Dingxu. 1994. The nature of Chinese wh-questions. *Natural Language & Linguistic Theory* 12:301-333.

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.

Tang, T. C. 1984. *Hanyu cifa jufa lunji* [Essays on Chinese Morphology and Syntax]. Student Books Co., Taipei, Taiwan.

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

Xu, L. and D. T. Langendoen. 1990. Topic structures in Chinese. *Language* 61 (1), 1-27.

Xu, L. (1996). Construction and destruction of theories by data: A case study. *CLS* 32:107-118.

Zhang, Ning. 2006. *The Minimal Syntax of Coordination*. National Chung Cheng University ms.

Zoerner, Cyril Edward. 1995. *Coordination: The syntax of &P*. University of California, Irvine PhD thesis.

**Notes**

[1] Examples are cited without the usual diacritics (*, ?, etc) because their acceptability is precisely what is at issue. Empty categories are indicated by *e*. ASP = aspect marker, CL = classifier, MOD = modifier marker.