LANGUAGE AND LINGUISTICS 13.3:437-470, 2012
2012-0-013-003-000049-1

Testing Adjunct and Conjunct Island Constraints in Chinese[*]

James Myers
National Chung Cheng University

A growing number of syntacticians are supplementing their own intuitions with formal experiments, collecting and analyzing acceptability judgments from theoretically naïve native speakers. This paper applies this experimental approach to test a set of interrelated hypotheses in Chinese syntax: that extraction from conjunct islands is more acceptable than extraction from adjunct islands; that extraction from adjunct islands truly violates grammar rather than merely affecting sentence processing; and that relativization involves movement while topicalization does not. The results support the first hypothesis but challenge the other two. The study also demonstrates how quick and simple formal judgment experimentation can be.

Key words: Chinese, syntax, island constraints, acceptability judgments, methodology

1. Introduction

Linguistics is often thought of as a branch of cognitive psychology, but for historical reasons, linguists and psychologists use dramatically different methodologies: psychologists use experiments, while theoretical linguists, syntacticians in particular, typically use their own native-speaker intuitions. Recently a growing number of linguists have argued that it is time for syntacticians to adopt the methods of their psycholinguistic colleagues, and test sentence acceptability using explicit experimental protocols (Schütze 1996, Cowart 1997, Featherston 2007, Myers 2009a, 2009b).

While this shift is a positive sign of the maturation of linguistics as a science, it seems counterproductive to oblige syntacticians to test all of their claims with such methods, given how time-consuming these methods can be, and how uncontroversial many syntactic judgments are (as opposed to their theoretical interpretation). A more

[*] This study began with a discussion with Niina Zhang, who helped with the material design and syntax, and translated the title and abstract, but is otherwise innocent of responsibility for the end result. Thanks also to Ko Yuguang, Yang Chentsung, and especially two anonymous reviewers for their thoughtful comments on both technical details and larger issues. The usual caveats apply even more forcefully than usual here, since there simply is no room in a short journal article to accommodate all of the heatedly debated views on linguistic methodology.

appropriate way to think about experimental syntax is that it offers syntacticians more flexibility in testing their empirical hypotheses (in addition to other data sources, including corpus analysis, acquisition, and neurolinguistics, none of which will be discussed in this paper). Traditional informal judgments may often suffice, while at other times judgments may be so delicate, that is, so sensitive to pragmatics, lexical content, or parsing constraints, that the best way to understand them is to control or systematically vary these variables in a full-fledged experiment. At still other times, however, the syntactician may choose to adopt experimental techniques that go beyond the traditional informal methods only slightly, just enough to make it possible to evaluate the statistical reliability of the claims. Myers (2009a) calls this approach, lying midway between traditional informal judgments and full-fledged experimentation, small-scale syntactic judgment experimentation.

This study describes a small-scale judgment experiment testing a set of hypotheses in Chinese syntax. The goal of the study is to combine the speed and ambition of traditional syntax, where many hypotheses are tested using relatively few sentences and native-speaker judges, with the empirical robustness of experimental psychology. Thus this study tests not just one hypothesis, but a network of them, in a single small experiment. These hypotheses concern adjunct islands, conjunct islands, the nature of Chinese topicalization as movement, and the distinction between grammatical competence and the performance of sentence parsing. Despite its ambitions, the experiment was so simple to design, run, and analyze that the entire process took only a day and a half, using no special-purpose tools.

Section 2 reviews the notion of small-scale syntactic judgment experiments. Section 3 turns to the theoretical focus of the study: the nature of adjunct islands, conjunct islands, and their interaction with topicalization and relativization. Section 4 describes the experiment itself, from design through results. These results support some commonly made claims in the literature, in particular the unacceptability of extracting from adjunct islands, and the apparent status of topicalization as movement. Yet they also challenge other traditional claims, finding that extracting from conjunct islands is relatively acceptable, and that the adjunct island constraint may be modulated by processing, not by grammar alone. Section 5 provides a brief conclusion.

2. Small-scale syntactic judgment experiments

Phillips & Lasnik (2003:61) are right to emphasize that the “[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.” Even when a syntactician tests his or her own judgments informally, the basic elements of a real psycholinguistic experiment are present: stimuli (the sentences), responses (the judgments), an experimental design (typically involving minimal pairs of sentences predicted to contrast in acceptability), and a well-defined task that is no more unnatural than the lexical decision task ubiquitous in experimental psycholinguistics. Crucially, the judgments are of acceptability, which is a pretheoretical feeling accessible even to naïve native speakers, not of grammaticality, which is the formal status of a sentence as defined by some grammatical theory. Thus even though they derive from intuition, acceptability judgments are potentially replicable, typically being shared among linguists who sharply disagree on their theoretical implications (Schütze 1996, Myers 2009a).

Moreover, the simplicity of the informal methods used by theoretical syntacticians to collect acceptability judgments should not count against them. The value of a scientific methodology should not be measured by its complexity (e.g. whether it is messy enough to require the scientist to wear a white lab coat), but by its results, and it is clear that informal judgments of native-speaker acceptability are a powerful tool for discriminating amongst theoretically interesting hypotheses. To take a simple example, any theory of syntax that ignores the island constraints of Ross (1967) will have considerable difficulty explaining the extreme unacceptability of a sentence like “Who did John believe the claim that Bill saw?” Such constraints may be argued to derive from pragmatics or language processing rather than autonomous syntax (e.g. Ambridge & Goldberg 2008), but the fact remains that the constraints do exist in some sense, and this fact is known to us primarily through informal acceptability judgments.

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) have puzzled over the implications of judgment “haziness” (p.290), and debates over judgments (whether they are factually correct, and if they are, whether they reflect syntactic competence rather than pragmatics or processing) are common at conferences, in peer reviews, and in the published literature (Schütze 1996, 2011, Myers 2009a).

In response, some syntacticians have recently begun to advocate experimental methods more in accord with psycholinguistics (Cowart 1997, Featherston 2007): multiple stimuli and participants (naïve speakers rather than the potentially bias-prone syntacticians themselves), factorial designs (crossing two or more factors to create lexically matched sentence sets, rather than testing sentences in isolation or, at best, in minimal pairs), gradient response measures (multivalued scales rather than the traditional binary yes/no judgment), filler items (theoretically irrelevant sentences use to hide the pattern being tested in the target items), randomization of presentation order (to avoid confounding syntactic factors with fatigue, practice, or cross-sentence influences), counterbalancing of sentence sets across participants (i.e. distributing sentences so that all participants see all sentence types but no participant gets more than one sentence from each lexically matched sentence set), and statistical analysis.

Judgments collected in this way have sometimes reconfirmed some widely accepted phenomena, such as *that*-trace effects in English (Cowart 1997). In a full-scale judgment experiment on naïve native speakers, Sprouse & Almeida (to appear) even managed to replicate virtually all of the hundreds of judgments in a standard syntax textbook (Adger 2003); Sprouse et al. (2011) report a similar replication rate for judgments in *Linguistic Inquiry* articles published between 2001 and 2010. Yet judgment experiments have also revealed hitherto unsuspected complexity. This includes the discovery of German *that*-trace effects (Featherston 2005a) of such subtlety that informal methods had failed to detect them, and the falsification by Clifton et al. (2006) of a key theoretical claim proposed in Kayne (1983). Moreover, syntactic claims relating to gradience (including the "haziness" noted by Chomsky 1981) or grammar/processing interactions cannot be tested at all without the appropriate experimental methods (Featherston 2007, Myers 2009a). For example, Ambridge & Goldberg (2008) were able to test their hypothesis that island constraints are related to constraints on the processing of information structure only because they tested many systematic sets of sentences on many naïve speakers, rather than merely judging sentences by themselves. For other examples and further discussion of the benefits of explicit experimental methods in syntax, see Featherston (2007), Myers (2009a), and Schütze (2011).

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 the *that*-trace experiment described by Cowart (1997) is compared with the theoretical analysis of the *that*-trace effect by Chomsky & Lasnik (1977): in the former, great effort is expended simply to establish the existence of *that*-trace effects in naïve speakers, whereas in the latter, such effects form just part of a wide range of observations in a much more ambitious analysis.

Given such considerations, Myers (2009a, 2009b) argues that what is needed is a middle ground between the current status quo of self-elicited informal judgments and full-scale laboratory experimentation. Methods that are powerful enough to yield statistically valid results can still be simple and cheap enough to apply quickly. This approach, which Myers (2009b) calls small-scale syntactic judgment experimentation, is adopted in the present study.

Table 1 lists the similarities and differences among traditional informal judgment methods (i.e. those that Phillips & Lasnik 2003 called "trivially simple"), full-scale psycholinguistic experimentation of the sort reviewed above, and the small-scale experimentation advocated here. Shading is used to highlight the features that small-scale experimentation shares with informal or full-scale methods.

Table 1: Features of small-scale syntactic judgment experimentation

Feature	Informal	Small-scale	Full-scale
experimental design	minimal pairs, if any	factorial	factorial
statistical analysis	no	yes	yes
number of sentence sets	may be as few as one[1]	multiple (a few)	multiple (many)
number of participants	may be as few as one	multiple (a few)	multiple (many)
presentation order	irrelevant	random	random
judgment scale	binary	binary	usually gradient
filler items	no	no	yes
counterbalancing	no	no	yes
type of participants	theoretician(s)	semi-naïve	totally naïve

The most important feature that small-scale experimentation shares with full-scale experimentation is the use of factorial designs, where syntactic variables are systematically crossed to create lexically matched sentence sets, rather than testing arbitrary lists of independent sentences. As noted in Myers (2009b), the factorial approach is already implicit in traditional syntactic methodology, where it is standard practice to compare the relative acceptability of minimal pairs of sentences. Yet as Myers goes on to show, syntacticians do not always adopt the appropriate experimental designs, and when they do so, they tend to neglect multi-factor designs, which are needed whenever syntactic factors interact with each other, as is very often the case. For example, the *that*-trace effect is defined by the interaction between *that* and the trace, not by each element alone, and thus testing it requires not merely minimal pairs, but quartets of sentences (*that* + subject gap, *that* + object gap, no *that* + subject gap, no *that* + object gap).

The statistical analysis of an experiment depends on its design to define the basis of comparison for calculating statistical significance. Without a basis for comparison, we cannot tell if the acceptability of any isolated sentence is particularly high or low relative to a judge's baseline response bias (e.g. perhaps the speaker tends to accept

[1] In traditional syntactic practice, informal judgments often come from speakers other than the researcher him- or herself (including conference attendees and reviewers), and often more than one sentence set is tested, even if only one appears in print. However, since methodological documentation is not part of traditional syntactic practice, the evidence for these de facto standards is necessarily anecdotal.

most sentences, regardless of their grammatical status). Statistical analysis also depends on testing multiple sentences on multiple speakers, the rule of thumb being the more data, the more reliable the analysis. However, for ease and speed of testing, small-scale experimentation exploits powerful statistical tools, described later in this paper, that require relatively few speakers or sentences.

Small-scale experimentation also adopts the randomized sentence presentation order of full-scale experimentation, to avoid confounding theoretically relevant syntactic factors with nuisance variables like fatigue, practice, and cross-sentence influences.

In other features, small-scale experimentation is more similar to traditional informal judgments. In particular, the judgment scale is the familiar binary yes/no contrast, between relative acceptability and relative unacceptability, which makes the task a quick job even for naïve participants. Despite the preference that psycholinguists generally have for gradient measurement scales (e.g. reaction times), and despite the emphasis given to gradient judgments in some of the experimental syntax literature (e.g. Bard et al. 1996, Featherston 2005b, 2007), recent research has found little advantage for them in studying syntactic judgments experimentally (Weskott & Fanselow 2009, 2011, Bader & Häussler 2010, Sprouse 2011). One reason for this is that even if acceptability is gradient in reality, this gradience will emerge when binary judgments are averaged across sentences or speakers (Cowart 1997). Another is that raw acceptability itself, even when measured on a gradient scale, is far less gradient than had previously been thought (Sprouse 2007): native speakers naturally prefer to judge sentences as "good" or "bad."[2]

Small-scale experimentation also eschews two full-scale techniques that are intended, among other purposes, to hide the experimental design from the participants, namely the use of irrelevant filler sentences and counterbalancing (i.e. splitting up lexically matched sentence sets across participants). Eliminating these techniques greatly simplifies the preparation of sentence lists. The lack of counterbalancing also increases statistical power, since all sentences are given to all participants, increasing the total number of data points.

[2] Another intuitive judgment method is the forced-choice task, where participants choose between two minimally different options. In judgment experiments, these are two sentences identical except for the syntactic property of interest. Unsurprisingly, this method is highly sensitive to syntactic differences between the competing sentences (Sprouse & Almeida 2011). No method is perfect, however. In particular, the forced-choice task seems to presume a methodological philosophy inconsistent with the one underlying the use of counterbalancing and fillers, described below, since it makes the theoretically relevant contrast fully explicit to the participants. Moreover, this task is even further removed from natural language use than the yes/no judgment task, which at least simulates one aspect of the common experience of hearing or reading a sentence that seems (perhaps only at first) to be anomalous.

Although the use of fillers and counterbalancing are emphasized by some advocates of full-scale experimentation (e.g. Cowart 1997), their absence need not bias the results. It is true that the lack of fillers may permit some participants to make guesses about what the experiment is trying to test, and the lack of counterbalancing may encourage some participants to judge sentences relative to their lexically matched counterparts rather than for their own sakes, but there may well be many other participants who do not pick up on these cues, and even those who do may not be influenced the same way (e.g. they may make incompatible guesses about the purpose of the experiment). The randomization of sentence presentation order should reduce any such influences still further by separating lexically matched sentences, making them harder for speakers to compare explicitly, and by canceling out any asymmetrical cross-sentence influences (if for one speaker sentence A influences the judgment of sentence B, for another speaker B will appear before A). Thus while the lack of fillers and counterbalancing in small-scale experiments may add noise, as judgments become influenced by extra-grammatical forces differing from speaker to speaker and sentence to sentence, noise is not the same as systematic bias. Noise merely reduces sensitivity, and being small-scale, small-scale experiments can never be as sensitive as full-scale ones anyway. In a paper advocating full-scale syntactic judgment experiments, Featherston (2007) does not even include fillers and counterbalancing in his list of "essential" and "desirable" features (p.282).

Finally, regarding the type of experimental participant, small-scale experimentation stands midway between informal methods and full-scale experimentation in that it requires only that the participants not be the theoretician him- or herself, to reduce the risk of potential experimenter bias. They should be naïve to the purpose of the experiment and the relevant literature, but banning linguists or linguistics students entirely would throw out perfectly normal, and readily available, native speakers. Differences do exist in the judgment patterns of non-linguists versus linguists, as Dąbrowska (2010) found when she specifically sought them out. Nevertheless, such differences seem quite subtle; as already noted, multiple studies on naïve speakers have replicated patterns claimed in the theoretical literature (e.g. Cowart 1997, Featherston 2007, Sprouse & Almeida 2011). Rejecting all linguists as prospective experimental participants also begs the question of how much linguistic training is enough to "corrupt" one's judgments. The study reported in the present paper was conducted on students during their first semester in a masters program in linguistics. They were not, strictly speaking, linguistically naïve, but at the time the experiment was conducted, they had not yet even reached the textbook's discussion of islands and movement (the focus of the experiment).

It is important to emphasize that the differences between informal and small-scale judgment experiments are literally quantitative, not qualitative. In statistical terms, 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, 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. In syntax this generally means matching syntactically irrelevant variables (e.g. lexical content) within sentence sets, and testing multiple sets and speakers to see what patterns are robust enough to emerge amidst the variation. Syntacticians are already aware of the importance of matching lexical content in minimal pairs and minimal sets, and they also understand the importance of testing more than one sentence or speaker, often considering multiple sentences before choosing the "clearest" examples to present in their papers, and double-checking judgments with their colleagues. Syntacticians even have informal statistical intuitions, worrying when judgments seem to vary too much across speakers or sentences. In small-scale experimentation, these familiar principles (factorial design, control, sampling, and statistical analysis) are merely applied more systematically.

As inevitably happens with attempts to stake out a middle ground, small-scale judgment experimentation faces challenges from both extremes. There is no well-defined middle ground in any case, since the features of informal and full-fledged experimentation could probably be combined fruitfully in ways other than as defined here; experimenters are under no obligation to obey any particular methodological dogma. The advantages claimed here for small-scale experimentation over traditional informal judgments ultimately lie not in any particular feature, but in the spelling out of an explicit methodology of some sort, so that one's peers have some way of evaluating (or even statistically quantifying) the reliability of one's claims. Cowart (1997) calls judgments that meet such criteria "objective," in that they turn private intuitions into empirically justified facts that must be accepted by all claimants in a syntactic debate (whatever their preferred theoretical interpretation might be).[3]

On the other hand, once one starts down the road towards greater objectivity in judgment collection and description, there is no scientific reason to stop at small-scale experimentation, only considerations of convenience. I have already argued that there is no great risk in using binary judgment scales or neglecting fillers or counterbalancing, but these claims have yet to be empirically tested. The conventions of full-fledged experimentation undeniably exist for sound reasons (fillers and counterbalancing, in

[3] Of course, syntacticians who rely on informal judgments already believe that such judgments are empirically well-justified by existing community standards. However, given the explosion of full-scale judgment experiments in the past fifteen years or so, it seems that these community standards are rapidly changing. History suggests that all sciences become increasingly quantitative and methodologically rigorous, and theoretical syntax seems to be conforming to this generalization as well.

particular, both have long traditions in psycholinguistics), so theoretical syntacticians who wish to go "all the way" are certainly encouraged to do so. Yet insisting on full-fledged experimentation too strongly seems likely to intimidate syntacticians without any formal experimental experience at all, and even experienced experimentalists may find it useful to run a quick-and-dirty small-scale test of judgments just a bit too fuzzy to resolve with informal methods.[4]

The principles of small-scale experimentation (see Myers 2009b for further details) are so simple that they can be automated to a great extent. This is the purpose of the free program MiniJudge (Myers 2007), the latest incarnation of which is MiniJudge 2.0 (available at www.ccunix.ccu.edu.tw/~lngproc/MiniJudgeJava2.htm). Like earlier versions, it guides the novice experimenter through the design, running, and statistical analysis of small-scale judgment experiments involving one or two binary factors. The newest version brings greater flexibility and ease of use than ever before. In the design stage, MiniJudge 2.0 uses a more intuitive interface for generalizing the basic sentence set to new sets. In the experiment running stage, it includes the options of presenting sentences to participants on paper, on the experimenter's computer (as in a lab setting), or on a dedicated web site (participants are emailed the link, and data are automatically emailed to the experimenter when each participant finishes). Finally, in the analysis stage, MiniJudge communicates with the statistical analysis program behind the scenes, so that the experimenter need only deal with the result summaries and graphs.

MiniJudge is merely a tool, of course, and no tool can or should replace human creativity and insight. It may even be argued that experimental design software poses the risk of becoming a crutch, hindering rather than fostering a deep understanding of experimental logic. Indeed, small-scale judgment experiments are readily achievable without any special-purpose software (as are full-scale experiments; see Cowart 1997, 2012, for tutorials). The real power of small-scale experimentation lies in the hands of the experimenter him- or herself, not any particular tool, as will be demonstrated through the step-by-step description of the methods used in the present study.

Before I describe these procedures, however, I first describe the theoretical motivation behind the experiment.

[4] A reviewer estimates that a small-scale judgment experiment would save no more than two hours over a full-fledged one. This claim is also in need of empirical confirmation, and if testing it causes readers of this paper to become highly efficient full-fledged experimentalists, I would have no objection.

3. Adjunct and conjunct extraction in Chinese

Based on informally made syntactic acceptability judgments in Chinese and English, Huang (1982) concluded that extraction from adjuncts is universally disallowed, a proposal formalized in the Condition on Extraction Domain (CED). One of his Chinese examples is shown in (1) (his (33), p.466), where topicalization out of the *yinwei* 'because' clause is claimed to induce unacceptability.

(1) Zhangsan$_i$, [Lisi [yinwei [wo meiyou qing _$_i$]] hen bugaoxing].[5]
Zhangsan Lisi because I not invite very unhappy
'Zhangsan$_i$, Lisi was very unhappy because I did not invite _$_i$.'

When applied to adjunct extraction, the CED is not controversial as an empirical generalization (Stepanov 2007 reviews evidence against its application in extraction from subjects, and there continues to be debate over how it is derived; see also Nunes & Uriagereka 2000). The status of conjunct islands is less clear, however. Extraction from coordinate structures is generally assumed to be governed by the Coordinate Structure Constraint (CSC) proposed by Ross (1967). As observed by Grosu (1973), the CSC is composed of two subconstraints. The Conjunct Constraint forbids conjuncts themselves from moving out of a coordinate structure, as illustrated in English by the unacceptability of (2a) (Ross 1967), while the Element Constraint forbids any element within a conjunct from moving out of that conjunct, as illustrated by the unacceptability of (2b) (Lakoff 1986).

(2) a. What sofa$_i$ will he put the chair between [some tables and _$_i$]?
b. What kind of herbs$_i$ did you [[eat _$_i$] and [drink beer]]?

Following Grosu (1973), Lakoff (1986), and Culicover & Jackendoff (1997), among others, Zhang (2009) claims that the Element Constraint of the CSC, illustrated by (2b), can be violated when the conjuncts are semantically or pragmatically related. In (2b) there is no relation, intrinsic or discourse-related, between eating herbs and drinking beer, whereas in (3a-b), the notions of eating herbs and not getting cancer are related via the discourse context, making these two sentences acceptable (Lakoff 1986).

[5] Linguistic examples in this paper are cited without the usual diacritics (*, ?, etc) because their acceptability status is precisely what is at issue. Silent positions are indicated by _. ASP = aspect marker, CL = classifier, MOD = modifier marker.

(3) a. What kind of cancer$_i$ can you [[eat herbs] and [not get _$_i$]]?
b. What kind of herbs$_i$ can you [[eat _$_i$] and [not get cancer]]?

Zhang (2009) therefore agrees with Lakoff (1986) that the Element Constraint of the CSC is not a syntactic constraint, but a semantic or pragmatic constraint. Elements can only be extracted from conjuncts in what Wälchli (2005) calls natural coordinate structures, like (3). With accidental coordination, as in (2b), extraction is not possible. Even beyond their role in explaining CSC violations, the contrast between natural and accidental coordination has morphological or syntactic effects in a wide variety of languages; Wälchli (2005) and Dalrymple & Nikolaeva (2006) collectively cite examples from English, German, Russian, Finnish, Georgian, Kurdish, Eastern Armenian, Lenakel, Tundra Nenets, Udihe, Erz'a-Mordvin, Aymara, and Babungo, among other languages.

A Chinese example of an Element Constraint violation is shown in (4) (Zhang 2009:137), where *baozhi* 'newspaper' can be coindexed with the gap internal to one of the conjuncts conjoined by the coordinator *erqie* 'and'. Zhang (2009) argues that the Conjunct Constraint, the other component of the CSC, can also be violated in natural coordination structures in Chinese (although not in English), but I will not examine this claim in this paper.

(4)	na	fen	[Baoyu	kan-le	_$_i$	erqie	hai	xie-le	biji]	de	baozhi$_i$
	that	CL	Baoyu	read-PRF		and	also	write-PRF	note	DE	newspaper

'the newspaper that Baoyu read and also wrote notes on it'

There are a number of good reasons to want reconfirmation of the empirical claims made about sentences like the apparently unacceptable CED violation in (1) and the apparently acceptable CSC violation in (4). These particular claims were made by linguists immersed in a tradition that both makes specific claims about the acceptability of such sentences and rewards the discovery of universals. Hence it is conceivable that judgments by Huang and Zhang about (1) and (4), respectively, are unconsciously swayed by theoretical bias. Moreover, the apparent fact that such a well-known constraint like the CSC can be circumvented raises the possibility that the CED may also be weaker than it is currently assumed to be.

Looking at these questions in Chinese is particularly interesting, since it is one of those "vast array of languages," alluded to in the quotation from Phillips & Lasnik (2003), about which most linguists (other than native speakers) know very little. Thus, unlike the case with English, Chinese linguists cannot count on their international readers being able to confirm the claimed judgments using their own intuitions. The most effective way to make empirical claims about Chinese convincing to a non-Chinese speaker would be to use "objective" methods, in the sense of Cowart (1997), as described in the previous section.

The need for objective Chinese judgments is particularly acute because Chinese judgments are notoriously controversial, perhaps more so even than in English. For example, Huang (1982) claims that in (5) (his (198), p.267), *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 & Pan (2003). Xu (1990, 1996) and Shi (1994) challenge other judgments in Huang (1982).

(5)	Ni	xiang-zhidao	shei	mai-le	shenme?
	you	wonder	who	buy-ASP	what

'What is the x such that you wonder who bought x?'

Some of the Chinese judgments in Aoun & 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). The difficulty of determining the scope of syntactic generalizations in Chinese is discussed by Xu (1996), who describes how Battistella & Xu (1990) found a strong preference for long-distance binding of the reflexive *ziji* in the same syntactic structures where Ho (1995) found a strong preference for local binding (Xu 1996 argues that the difference relates to the pragmatics of different verb types). Finally, although Myers (2009b) did confirm several of the claims made in Li (1998) in a small-scale judgment experiment with naïve Chinese speakers, some of Li's judgments were not replicated.

How could we go about designing an objective test of the CED and CSC in Chinese? The most fundamental prediction is that if Huang (1982) and Zhang (2009) are both correct, CED violations like (1) should remain significantly worse than CSC violations like (4) even when we test theoretically unbiased native speakers in sufficient numbers to permit statistical analysis. Yet before we rush to do this, however, we must first deal with the fact that these two specific sentences differ in many ways besides the adjunct vs. conjunct contrast. Judgment differences between them may thus reflect theoretically irrelevant variables like word frequency, pragmatic plausibility, or aspects of syntactic structure that have nothing to do with adjunct or conjunct islands per se.

One such syntactic aspect is particularly worrisome: sentence (1) uses topicalization to extract from the adjunct, while sentence (4) uses relativization to extract from the conjunct. Some Chinese topics seem to be base-generated, not derived through movement, as in the example in (6) (Xu & Langendoen 1985:19).

(6) Shuiguo, ta zui xihuan pingguo.
fruit he most like apple
'As for fruit, he likes apples most.'

Moreover, whether or not object gaps are base-generated or derived via ellipsis of one sort or another (Huang 1987, Aoun & Li 2008), Chinese does permit them, as shown in (7).

(7) Wo xihuan __.
I like
'I like it.' [in the appropriate discourse context]

Building on such observations, Xu & Langendoen (1985) have argued that all Chinese topics are base-generated, since even those that seem to involve movement can be analyzed instead as object gaps coindexed with base-generated topics. They therefore predict that Chinese can appear to violate island constraints through this non-movement-derived coindexing. An example is shown in (8) (their (63c), p.15), in which a topic is coindexed with a gap within a relative clause. Although this structure would violate the complex NP constraint (Ross 1967) if topicalization involved movement, they note that "our observations of a considerable number of native speakers reveal that the occurrence of such structures is by no means rare" (p.15), consistent with a non-movement-derived analysis. The judgments are not secure, however, with Xu & Langendoen admitting that "[s]ome readers may question the acceptability of sentences" like (8) (p.15).

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

If Xu & Langendoen are correct about topicalization not being movement, the CED would be irrelevant for (1), since it would not involve extraction, and if it is unacceptable, it must be for some other reason. By contrast, regarding putative movement from the relative clause crucial to the CSC violation in (4), it appears that nobody has explicitly argued against such an analysis: Chinese syntacticians seem to be in agreement that movement is truly necessary here.

To tease apart this complex set of issues, we should test sentences varying only in the binary factors [±Adjunct] (adjunct islands vs. conjunct islands) and [±Topic] (topicalization vs. relativization), while keeping lexical content and other aspects of syntactic structure as constant as possible. This is done in the quartet of sentences in (9a-d). Note that in these sentences the target clauses are always embedded, to make it unambiguous that the topics are sentence-initial (and hence truly are topics). In all of these sentences, constituents (adjuncts [+Adjunct] or conjuncts [–Adjunct]) contain gaps that are coindexed with elements external to these constituents (via topicalization [+Topic] or relativization [–Topic]). Thus all of them contain island violations of one sort or another. Moreover, sentences (9c-d), which involve violations of the Element Constraint of the CSC, involve natural coordination, to control for semantic and pragmatic influences.

(9) a. [+Adj, +Top] Na-fen zuoye$_i$, Lisi shuo ta [ruguo xie-le _$_i$],
that-CL homework Lisi say he if write-ASP
jiu kan baozhi.
then read newspaper
'That homework$_i$, Lisi said [if he wrote _$_i$] then he'll read the newspaper.'

b. [+Adj, –Top] Na-fen Lisi shuo ta [ruguo xie-le _$_i$] jiu kan
that-CL Lisi say he if write-asp then read
baozhi de zuoye$_i$, jiu zai nali.
newspaper MOD homework just is there
'That homework$_i$ that Lisi said [if he wrote _$_i$] then he'll read the newspaper is there.'

c. [–Adj, +Top] Na-fen zuoye$_i$, Lisi shuo ta [xian xie-le _$_i$],
that-CL homework Lisi say he first write-ASP
ranhou kan-le baozhi.
and then read-ASP newspaper
'That homework$_i$, Lisi said he [first wrote _$_i$] and then read the newspaper.'

d. [–Adj, –Top] Na-fen Lisi shuo ta [xian xie-le _$_i$] ranhou
that-CL Lisi say he first write-ASP and then
kan-le baozhi de zuoye$_i$, jiu zai nali.
read-ASP newspaper MOD homework just is there
'That homework$_i$ that Lisi said he [first wrote _$_i$] and then read the newspaper is there.'

If Huang (1982) is right about the unacceptability of CED violations, and Zhang (2009) is right about the acceptability of violations of the Element Constraint of the CSC in natural coordination, we then predict that [+Adjunct] sentences like (9a-b) should be worse, on average, than [–Adjunct] sentences like (9c-d). If Xu & Langendoen (1985) are right, and topics (but not relative clauses) are base-generated, then we predict an interaction between the [Adjunct] and [Topic] factors: [+Adjunct, –Topic] sentences like (9b) should be worse than [–Adjunct, –Topic] sentences like (9d), but [+Topic] sentences like (9a,c) should not differ. That is, in sentences with topicalization, there should be little or no adjunct island effect (so that the average unacceptability of (9a-b) noted above would actually be due solely to the unacceptability of (9b)). These predictions can be schematized as in Figure 1, which shows the relative degree of acceptability expected for the four types of structures.

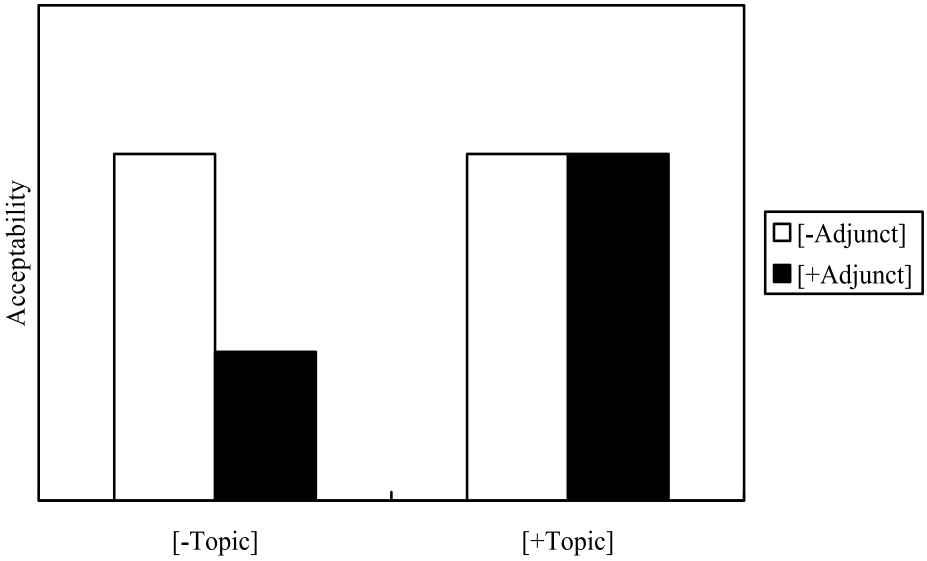

Figure 1: Expected acceptability pattern

Although these sentences are well-matched and controlled, the complexity of the structures needed to accomplish the above goals seem likely to make them far more difficult to parse than the original (1) and (4). This suggests that traditional informal methods of judgment collection will not work, since individual judgments will probably not be very sharp. We may only be able to see a pattern if we aggregate judgments from a number of speakers on a variety of sentences.

Not only are all of the sentences in (9) somewhat complex, but they are not fully matched in parsing difficulty either. Chinese relative clauses as in (9b,d) create center-embedded structures, which, as is well known, are difficult to parse despite being grammatical (Chomsky 1965). A typical example in English is shown in (10c), which can be seen to be grammatical via its relationship with the sentences in (10a-b). The grammaticality of (10c) can also be demonstrated by the fact that (10d) has the same center-embedded structure but is much easier to parse than (10c), in part because the more deeply embedded portions are progressively shorter. Thus the lower acceptability of (10c) relative to the other sentences must relate to sentence processing, not grammar.[6]

(10) a. The mouse ran.
b. The mouse the cat chased ran.
c. The mouse the cat the dog bit chased ran.
d. The two frightened mice an old cat I hate chased ran far away.

In contrast to the right-branching structure of (9c), repeated below in (11a), (9d) shows center-embedding, as highlighted in (11b). If Zhang (2009) is correct, both of these CSC extraction structures should be grammatical, but even if they are, the acceptability of (9d)/(11b) may be reduced through parsing difficulties.

(11) a. Na-fen zuoye$_i$, [Lisi shuo [ta xian xie-le _$_i$, ranhou
that-CL homework Lisi say he first write-ASP and then
kan-le baozhi]].
read-ASP newspaper
'That homework$_i$, Lisi said he first wrote _$_i$ and then read the newspaper.'

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

Such a parsing effect could be detected in my experiment if sentences with relative clauses ([–Topic]) are judged less acceptable than sentences with topicalization ([+Topic]), whether or not the sentences have adjunct or conjunct islands. This pattern could not be

[6] The argument here is entirely standard, though as a reviewer points out, it makes the (also entirely standard) assumption that grammar does not care about lexical differences like those between (10c) and (10d). Contra another reviewer, however, the argument does not depend on the assumption that grammars cannot count, since (10d) has just as many embeddings as (10c), yet is clearly more acceptable.

due to grammar alone, since relativization and topicalization are both allowed by Chinese grammar.

Another independent way to distinguish grammar from processing would be to exploit the phenomenon of syntactic satiation, whereby judges presented with a series of ungrammatical sentences find them increasingly more acceptable (Snyder 2000). Grammatical knowledge should be quite stable over the course of an experiment, but processing, by its very nature, fluctuates considerably; the processing of one sentence readily exerts influences on the processing of later ones (Luka & Barsalou 2005). This phenomenon is normally considered a nuisance; Snyder (2000:575) notes that syntactic satiation is also called "linguists' disease," dulling native-speaker intuitions over the course of a syntactic career. Yet as Snyder (2000) also suggests, it can potentially be a useful diagnostic tool as well, since the observation that intuitions can shift suggests that satiation "reflects limitations on sentence processing" rather than competence (p.580). Taking this as a strict principle (something Snyder himself declined to do), satiation of the [±Topic] contrast would further support the notion that any effects of this factor involve processing, rather than grammar per se.

Since Snyder's pioneering work, Sprouse (2009) has cast some doubt on the replicability of satiation, observing that some studies find it and others do not, depending on the type of grammatical violation and the task. He proposes that satiation is caused by speakers attempting to balance the number of "yes" and "no" responses in a syntactic judgment experiment, so if most of the sentences in the experiment are ungrammatical, participants will attempt to counter their earlier "no" responses by increasing their "yes" responses later on. However, as noted by Ko (2007), this explanation for satiation fails to explain why acceptability judgment shifts can also happen with gradient judgment scales, not just binary yes/no scales, and can involve a reduction, not just an increase, in acceptability (anti-satiation). Luka & Barsalou (2005) also found that acceptability can increase even for grammatical sentences.

Thus Snyder's more general suggestion to add (anti-)satiation effects to the syntactician's arsenal of diagnostic tools remains worthy of exploration, given how easy they are to measure even in a small-scale judgment experiment.[7] Statistically, change in

[7] A reviewer challenges the theoretical importance of satiation, rightly noting that satiation by itself cannot index extra-grammatical influences (especially since there is still considerable controversy about the nature, or even conceptual coherence, of the putative grammar/processing dichotomy), that nobody knows what cognitive mechanisms underlie satiation, that satiation does not occur with long-distance dependencies despite their heavier use of processing resources, and that satiation experiments have yielded inconsistent results, as just acknowledged here. Yet the main point stands: in a small-scale experiment satiation can be measured for free, so why ignore potentially interesting data *a priori*? See also Goodall (2011).

the strength of a factor over the course of an experiment can be measured as an interaction between the factor and the presentation order of the sentences, in which the contrast between the two values of the factor (e.g. [+Adjunct] vs. [–Adjunct]) is weaker for later-presented sentences (Myers 2007, 2009b). In this regard, it seems relevant that in her judgment experiments, Hiramatsu (2000) found no evidence of adjunct island satiation in English, in contrast to other island violations that did satiate. In Chinese, Ko (2007) did find shifting judgments with adjunct island violations, but the shift involved anti-satiation.[8] He suggests that judging a series of complex sentences of various sorts (temporarily) improved parsing ability, which then made it easier for participants to recognize adjunct island violations as truly unacceptable. Put together, these observations suggest that the CED may be a particularly robust grammatical constraint, not subject to fluctuations in processing.

In short, the complexity of the sentences in (9) blurs judgments about them, and satiation can only go beyond the nuisance it is in isolated judgments if it is systematically examined in tests of multiple sentences. Thus the "trivially simple" methods defended by Phillips & Lasnik (2003) are simply inadequate here.

4. Testing adjunct and conjunct islands

The theoretical goal of the present study was to provide a novel test of the hypotheses introduced in the previous section, namely that CED violations are ungrammatical (i.e. violate the grammar) while CSC violations are not (i.e. the CSC is not a grammatical constraint but a semantic/pragmatic constraint), that relativization involves movement but topicalization does not, and that center-embedded structures lower acceptability by making parsing more difficult. These hypotheses predict, respectively, that acceptability should be lower for [+Adjunct] than for [–Adjunct] sentences, that the [Adjunct] and [Topic] factors should interact such that [+Adjunct] sentences are worse than [–Adjunct] sentences only in [–Topic] (relativized) structures, and that [–Topic] sentences should be less acceptable than [+Topic] sentences. Moreover, by affecting parsing rather than grammar, the factor [Topic] may be expected to show satiation, that is, a weakening of the contrast between [+Topic] and [–Topic] sentences over the course

[8] In her first satiation experiment, Hiramatsu (2000) seems to define satiation solely in terms of an increase of "yes" responses (if she followed the method she describes on p.97 for an earlier study), so it is conceivable that the number of "no" responses may have increased instead (anti-satiation). However, it seems unlikely that such a clear pattern would not have been reported at all. Moreover, in her second experiment, a graph giving a direct comparison between no-to-yes and yes-to-no judgment changes (p.167) makes it clear that adjunct island violations showed no anti-satiation.

of the experiment, whereas the factor [Adjunct], reflecting the grammatical CED constraint, should show no satiation.

The methodological goal of the experiment was to test all of these hypotheses as simply as possible, extending the traditional method just enough to permit statistical analysis, consistent with the principles of the small-scale judgment experiment, without using any special-purpose tools.

4.1 Methods

4.1.1 Participants

The participants were twenty Chinese native speakers, graduate students in my linguistics program in Taiwan. They were familiar with the notion of acceptability judgments and basic syntactic theory but knew nothing about the theoretical issues examined in this experiment. The students were repaid by teaching them about the experiment's goals, methods, and results.

4.1.2 Design and materials

As noted above, the two factors in the experiment were [Adjunct] (whether the sentences contained adjunct islands or conjunct islands) and [Topic] (whether the sentence structures involved topicalization or relativization). To generate the additional sets of sentences needed for statistical analysis, the sentences in (9) 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 *Lisi* (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 (= 4 × 2 × 4) sentences, a very short list by psycholinguistic standards. The full list of materials is given in the appendix.

Consistent with the small-scale nature of this study, neither counterbalancing nor fillers were used: all participants received all and only these 32 experimental items. Presentation order was randomized separately for each survey form.

Creation of the survey forms was done semi-automatically in Microsoft Excel following the step-by-step guide in Cowart (1997), and took only an hour or two. The advantage of using a software tool, rather than creating all of the sentences by hand, is that it ensures that the intended syntactic structures are consistent regardless of lexical content. The procedure involved first writing sentence components in cells at the left side of the spreadsheet, then entering Excel functions into cells at the right that would copy these components in accordance with the [Adjunct] and [Topic] factors. For example, to automatically topicalize or relativize, Excel functions were distributed in cells as illustrated in Figure 2, where row 1 shows a topicalized structure and row 2 shows a relativized structure.

	A	B	C	D	E	F	G	H	I	J	K	...	L	M
1	那份	那份	李四說他	寫了	作業	看	報紙	=A1	=G1	，	=C1	...		
2	那份	那份	李四說他	寫了	作業	看	報紙	=A2			=C2	...	的	=G2

... realized as ...

	A	B	C	D	E	F	G	H	I	J	K	...	L	M
1	那份	那份	李四說他	寫了	作業	看	報紙	那份	報紙	，	李四說他	...		
2	那份	那份	李四說他	寫了	作業	看	報紙	那份			李四說他	...	的	報紙

Figure 2: Schematic distribution of cell functions in Excel

The sentence components were then combined within each row using Excel’s concatenation function (e.g. =H1&I1&J1&K1&L1&M1, for the first row in Figure 2). Once the master list of 32 sentences had been created, copies were pasted into separate sheets, one per survey. The individual random orders were created by entering Excel’s random number function (=RAND()) into the first column, before each sentence, and then sorting the rows according to the first column.

Presentation order was also a component of the experimental design, and was treated as a factor in its own right in the statistical analysis. There are two reasons for doing this. First, factoring order effects out in the statistics increases our confidence still further that the main effects are truly due to the experimentally important factors, and not artifacts of the survey-completion process itself. Second, by looking at the interaction between order and the factors, we can see how the influence of these factors on judgments changed over the course of the experiment. If judgments for putatively ungrammatical structures become more positive, this would represent a case of syntactic satiation.

4.1.3 Procedure

Each survey, with all sentences listed together on a single piece of paper, asked for binary yes/no judgments. Instructions were given orally. Because the participants were linguistics students familiar with the notion of acceptability judgments, no tips were given on what was meant by this notion (e.g. they received no examples of acceptable and unacceptable sentences to use as a basis of comparison). They were encouraged to go with their first reaction; if participants were not sure of their judgments, they were to make their best guess. Asking for clarification during the task was not allowed, nor was skipping sentences, whether or not the participant intended to return to them later.

Surveys took participants about ten minutes to complete, either as an in-class exercise or soon afterwards.

4.2 Results

The responses from 20 participants to the 32 sentences, with none skipped, generated 640 data points. 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).[9]

The results were analyzed using the free statistics program R (R Development Core Team 2011), available for download from www.r-project.org. Due to its power, flexibility, frequent updates, ready availability, and lack of cost, this program has become quite popular in quantitative linguistics (Baayen 2008, Johnson 2008, Gries 2009).

The key results are illustrated in Figure 3, which shows the proportion of "yes" judgments across participants as a function of [Adjunct] and [Topic]. These proportions were calculated by importing the tab-delimited results file into Excel, and computing the mean values of the 0s and 1s representing individual judgments using the Excel cell function =AVERAGE().

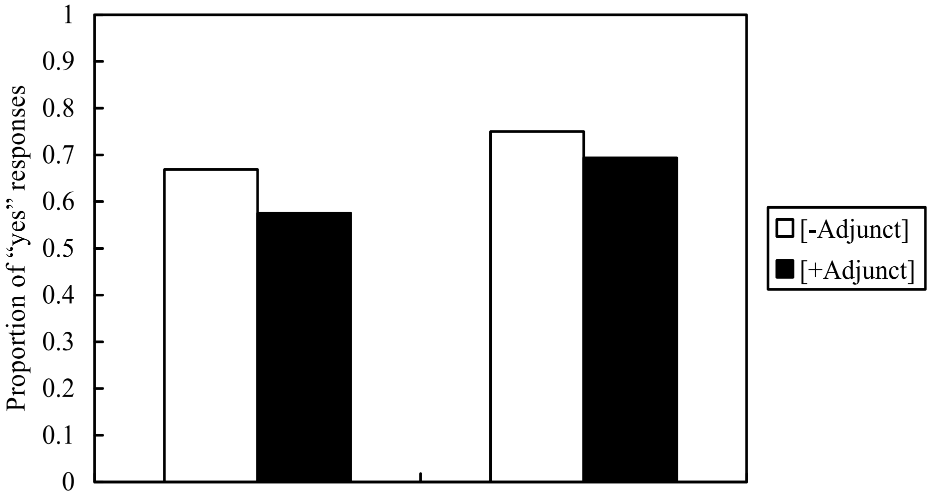

Figure 3: Main judgment results

[9] The data file synexp.txt is available at http://www.ccunix.ccu.edu.tw/~lngmyers/synexp.txt.

The graph suggests that sentences with adjunct extraction (black bars) were less likely to be judged as acceptable than sentences with conjunct extraction (white bars), whether or not they involved relativized structures ([–Topic]) or topicalized structures ([+Topic]).

Because the responses were binary yes/no judgments, they were analyzed with a generalized form of logistic regression, the statistical heart of the VARBRUL family of programs widely used in sociolinguistics (Mendoza-Denton et al. 2003). Ordinary logistic regression assumes that all of the data are independent of each other, an assumption that is violated here: each participant supplied 32 data points, in accordance with his or her own systematic preferences. In order to take the fixed factors of [Adjunct] and [Topic] into account at the same time as the random factors of participants and sentences, I used mixed-effects logistic regression, which combines fixed and random variables into a single equation (Baayen 2008, Myers 2009b). This analysis was conducted using the lme4 package in R (Bates et al. 2011); mixed-effects logistic regression can also be run in the commonly used statistical programs SAS and SPSS.

For the benefit of readers who may wish to run a similar study of their own, I present the R commands exactly as I used them, though of course commands in other statistical programs will differ in detail. The first set of commands were those in (12), which installed the lme4 package (which only needs to be done once), activated it for this session, and loaded the data from the tab-delimited text file. Variables are underlined, to distinguish them from built-in R command terms.

```
(12)  install.packages("lme4")
      library(lme4)
      synexp = read.table("synexp.txt", header=T)
```

Since mixed-effects logistic regression is a form of regression, the numerical variable of sentence order could be treated as just another factor. This factor, called Order, consisted of an integer for each participant for each sentence, representing what position in the survey list that participant was presented with that sentence (first position was indicated with 1, and so on). My analysis used a full model containing all three factors (Order, [Adjunct], [Topic]) and all possible (two-way and three-way) interactions. Thus the regression equations had the schematic structure shown in (13) (the reference to the binomial family indicates that this is logistic regression, not ordinary linear regression).

```
(13)  Judgment ~ Order * Topic * Adjunct, family = "binomial"
```

By treating both fixed and random variables in the same regression equation, mixed-effects models make it possible to test which of the random variables actually contribute to the description of the data. This is done by creating two mixed-effects models, one with only speakers as random variable and the other with both speakers and sentences, as in (14a-b).[10] I then ran a likelihood ratio test, a standard method for comparing related regression equations, using the R command in (14c). Only if it yields a statistically significant result ($p < .05$) should the more complex model be preferred (see also Myers 2007 for how this is done automatically in MiniJudge).

```
(14)  a.  model.s = lmer(Judgment ~ Order * Topic * Adjunct
                + (1|Speaker), data = synexp, family = "binomial")
      b.  model.ss = lmer(Judgment ~ Order * Topic * Adjunct
                + (1|Speaker) + (1|Sentence), data = synexp, family = "binomial")
      c.  anova(model.s, model.ss)
```

In this case, the more complex model in (14b), taking both cross-speaker and cross-sentence variability into account, was indeed statistically superior to the simpler one in (14a), yielding a *p* value below .05.[11] The core results of this analysis were generated with the R command in (15), which also rounds the values to four post-decimal digits.

```
(15)  round(summary(model.ss)@coefs,4)
```

The output of the command in (15) is shown in Table 2. Each row represents a potential predictor of the relative probability of giving a “yes” or “no” response. R’s notation “X:Y” refers to the interaction between the factors X and Y, that is, how the effect of X is modulated by Y or vice versa. The intercept is the overall bias to respond “yes” or “no” even after all other factors are taken into account. The columns represent four statistical values associated with each predictor. The estimate reflects the weight of the predictor; positive estimates indicate that the predictor is associated with more “yes” than “no” responses, while negative estimates indicate the reverse. Standard errors and *z*

[10] Note that these models are inspired by the traditional approach to experimental analysis in psycholinguistics (crossing all fixed factors but not crossing them with the random variables). Motivated readers can learn about more complex models for linguistic experiments in Baayen (2008), Johnson (2008), and Gries (2009).

[11] The *p* value indicates the probability that an estimate could be as far from zero as actually observed, in either direction, by chance alone. By the usual convention, $p < .05$ (less than a 5% chance) is considered to be so unlikely that the estimate can be taken as reflecting a genuine pattern.

values reflect the distance of the estimate from zero relative to the overall distribution of estimates. “Pr(>|z|)” represents the *p* value.

Table 2: Results of the statistical analysis

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3993	0.3691	1.082	0.2792
Order	0.0408	0.0112	3.6252	0.0003
Topic	0.3824	0.3002	1.2739	0.2027
Adjunct	-0.6355	0.2994	-2.1229	0.0338
Order:Topic	-0.0017	0.0114	-0.1527	0.8786
Order:Adjunct	0.0218	0.0114	1.9138	0.0556
Topic:Adjunct	0.3225	0.2987	1.0797	0.2803
Order:Topic:Adjunct	-0.0183	0.0113	-1.6139	0.1066

Looking first at the syntactic variables, the analysis revealed a statistically significant negative effect of [Adjunct] on judgments (*p* = .03326 < .05). Note the negative sign of the estimate for this factor, consistent with Figure 3, where the black bars (representing [+Adjunct] sentences) are lower than the white bars. The effect of [Topic] was not significant (*p* > .2). The interaction between [Topic] and [Adjunct] was also not significant (*p* > .2).

Turning now to Order, by itself this factor had a significant positive effect: overall acceptability increased over the course of the experiment. Theoretically more important, however, were the interactions with Order, which test for changes in factor strength, like satiation. The only such interaction with even a hint of statistical reliability was that between Order and [Adjunct], which just missed significance at *p* = .0556.

Since interactions are difficult to understand from regression values alone, I plotted the interaction between Order and [Adjunct] factor to examine it visually, as shown in Figure 4.

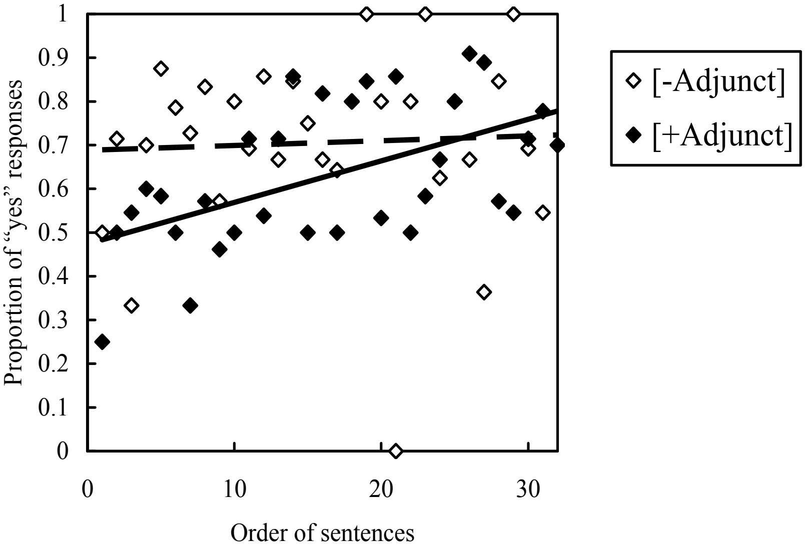

Figure 4: The influence of sentence order on [Adjunct] effects

This graph was created in Microsoft Word using values calculated in R. The reason for this two-step procedure is that R permits much greater flexibility in calculations than Microsoft Office, but due to this very flexibility, its graphing functions are less intuitive than those in Office. Here we needed to calculate the mean response (0 vs. 1) across surveys (participants) for all [–Adjunct] sentences that happened to appear first in the randomly ordered list, and then for all [–Adjunct] sentences that appeared second, and so on, and likewise for [+Adjunct] sentences. Since the orders were randomized, the number of sentences with each of these combinations of properties differed for each order position and [Adjunct] type, making the procedure cumbersome (though possible) to compute using cell functions in Microsoft Office. In R, however, the single command in (16) carried out this job automatically.

```
(16)  tapply(synexp$Judgment, list(synexp$Order, synexp$Adjunct), mean)
```

The matrix of values produced by this R command was then entered into the graphing tool of Microsoft Word to produce Figure 4. The dashed line shows the linear trend line for [–Adjunct] sentences, while the solid line shows the linear trend line for [+Adjunct] sentences. This graph makes it clear that the acceptability of the [–Adjunct] sentences remained constant throughout the experiment, while the acceptability of [+Adjunct] sentences increased.

4.3 Discussion

This single simple experiment was able to test all of the hypotheses raised in §3. First, CED violations ([+Adjunct]) were worse, overall, than CSC violations ([–Adjunct]), as jointly predicted by Huang (1982) and Zhang (2009).

Second, there was no significant interaction between the factors of [Adjunct] and [Topic], and in particular, the judged difference between [+Adjunct] and [–Adjunct] sentences was not significantly stronger for [–Topic] (relativized) sentences than for [+Topic] (topicalized) sentences. In other words, there was no support for the prediction of Xu & Langendoen (1985) that (adjunct) island effects should be weaker with topicalized structures, which they claim do not involve movement, than with relativization structures, which presumably do involve movement. As a null result, this does not actually falsify Xu & Langendoen's claim, but if future experiments continue to show no support for it, especially large-scale experiments of greater sensitivity, their claim becomes ever more doubtful.

Third, the factor [Topic] did not have a significant effect: center-embedded (relativized) structures were not judged significantly worse than right-branching (topicalized) structures. Given this, it is unsurprising that [Topic] also failed to show satiation (i.e. there was no interaction with sentence order indicating a weakening of this factor over the course of the experiment). Like the previous null result, this finding cannot be interpreted in isolation. One possibility is that given the length and complexity of all of the test sentences relative to those encountered in ordinary life, center-embedded structures may not have stood out as especially difficult. Of course, it is also possible that the small scale of this experiment simply made it too insensitive to detect a pattern that was actually present.

Fourth, there was a significant overall increase in the number of "yes" responses, but as shown by the marginally significant interaction between Order and Adjunct and Figure 4, this pattern was restricted to [+Adjunct] sentences. That is, while the judged acceptability of the sentences with CSC violations ([–Adjunct]) remained constant, the acceptability of sentences with CED violations ([+Adjunct]) increased. Thus the adjunct island constraint in Chinese had a tendency to satiate. If such a rapid change in acceptability over time is a mark of extra-grammatical forces at work, we might take this to mean that the CED has a processing component, at least in how it interacts with acceptability. This finding differs from that of Hiramatsu (2000), who found no satiation of the adjunct island constraint in English, and that of Ko (2007), who found that adjunct island violations in Chinese showed anti-satiation, becoming worse over the course of the experiment. As Sprouse (2009) noted, satiation is a delicate phenomenon. Nevertheless, we must still reject his hypothesis that satiation is always caused by judges attempting to balance the number of “yes” and “no” responses. Note that all four bars in Figure 2 represent proportions of “yes” responses over 50%, a trend reflected in the positive (albeit nonsignificant) estimate for the intercept in Table 2. Thus by increasing the number of “yes” responses for [+Adjunct] sentences over the course of the experiment, the speakers were skewing the yes/no distribution rather than balancing it out. The true cause of satiation (and anti-satiation), and the evidence it may provide about the nature of syntactic constraints, remains to be discovered.[12]

The methodological implications of the study were equally important. The ability of this study to test so many hypotheses simultaneously depended on the fact that these hypotheses fit so elegantly into a factorial design. Each of the eight rows in Table 2 represents a separate hypothesis, all tested in a single statistical analysis of one experiment. Syntactic factors do not always fit together so neatly, of course; even a single hypothesis may require multiple experiments to test, if the claim involves the unification of apparently disparate phenomena. In any case, the present experiment, despite being very easy to run, generated 640 data points (20 speakers × 32 sentences), quite a large amount of information. By using a mixed-effects regression technique, all of these data points could be analyzed, greatly increasing statistical power in comparison with a more traditional analysis of variance (ANOVA), which would require first averaging measurements by participant and by sentence. Such prior averaging would have left only 80 (= 20 speakers × 4 sentence types) data points for the by-participant analysis and only 32 (= 32 sentences) for the by-sentence analysis, for a total of only 112 data points, far fewer than the 640 analyzed here.

Preparing the experiment was also a quick job, using only familiar programs in Microsoft Office (Excel and Word). By deciding to include beginning linguistics students in the experiment, who could be tested as part of a class exercise, there was no need to spend the effort and cost to recruit fully naïve native speakers from the larger university population. The students tested here were naïve to the theoretical purposes of the experiment, which sufficed. Even in the highly unlikely event that some were aware of the CED and CSC, their judgment patterns did not reflect how these constraints are usually described in the literature: both should have given rise to equal degrees of unacceptability, and certainly satiation is not something that could easily be simulated on purpose. Of course, replication in fully naïve speakers would establish these points more firmly.

While no special tools were used to carry out this experiment, it would have been even simpler and quicker with the help of MiniJudge. All of the steps described above, from the creation of matched sentence sets through the creation and distribution of

[12] See footnote 7 for further caveats, however.

surveys to the statistical analysis of the results, can be done automatically with MiniJudge. The statistical analysis is still done with R, but in the latest version, MiniJudge 2.0, communication with R is handled by MiniJudge itself, so there is no need for the experimenter to deal with R commands. MiniJudge even draws the graphs automatically, of both types shown in Figures 3 and 4.

5. Conclusions

The entire process described in this paper, from initial conception of the experiment through experimental design and distribution and collection of the judgment surveys to initial statistical analysis, took a day and a half. Nevertheless, this study was still powerful enough to provide justifiable, objective 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; Chinese topicalization involves movement; acceptability for CED violations can shift.

Like all experiment reports, this paper describes a specific event, so further testing, particularly by skeptics of my conclusions, would be most welcome. Now that I have made my claims using a small-scale experiment, however, a convincing response will also have to involve a well-designed and conducted experiment, not merely informal judgments produced by the challenger him- or herself. In short, I hope that this experiment in experimental methodology has demonstrated the benefits of upgrading from the “trivially simple” traditional methods of syntactic judgment collection to the merely simple.

Appendix

Materials used in the small-scale syntactic judgment experiment. The original proper names used in these sentences, which for student interest were those of departmental colleagues, have been suppressed here.

	Adjunct	Topic	Connector	Sentence
1	+	+	因爲	那份作業，X說她因爲寫了，所以就看了報紙。
2	+	–	因爲	那份X說她因爲寫了所以就看了報紙的作業，就在那裡。
3	+	+	如果	那份作業，X說她如果寫了，就看報紙。
4	+	–	如果	那份X說她如果寫了就看報紙的作業，就在那裡。
5	–	+	然後	那份作業，X說她先寫了，然後看了報紙。
6	–	–	然後	那份X說她先寫了然後看了報紙的作業，就在那裡。
7	–	+	而且	那份作業，X說她不但寫了，而且看了報紙。
8	–	–	而且	那份X說她不但寫了而且看了報紙的作業，就在那裡。
9	+	+	因爲	那本小說，Y說他因爲看了，所以就寫了文章。
10	+	–	因爲	那本Y說他因爲看了所以就寫了文章的小說，就在那裡。
11	+	+	如果	那本小說，Y說他如果看了，就寫文章。
12	+	–	如果	那本Y說他如果看了就寫文章的小說，就在那裡。
13	–	+	然後	那本小說，Y說他先看了，然後寫了文章。
14	–	–	然後	那本Y說他先看了然後寫了文章的小說，就在那裡。
15	–	+	而且	那本小說，Y說他不但看了，而且寫了文章。
16	–	–	而且	那本Y說他不但看了而且寫了文章的小說，就在那裡。
17	+	+	因爲	那條裙子，所長說她因爲買了，所以就拍了相片。
18	+	–	因爲	那條所長說她因爲買了所以就拍了相片的裙子，就在那裡。
19	+	+	如果	那條裙子，所長說她如果買了，就拍相片。
20	+	–	如果	那條所長說她如果買了就拍相片的裙子，就在那裡。
21	–	+	然後	那條裙子，所長說她先買了，然後拍了相片。
22	–	–	然後	那條所長說她先買了然後拍了相片的裙子，就在那裡。
23	–	+	而且	那條裙子，所長說她不但買了，而且拍了相片。
24	–	–	而且	那條所長說她不但買了而且拍了相片的裙子，就在那裡。
25	+	+	因爲	那個皮包，Z說她因爲偷了，所以就拿走了錢。
26	+	–	因爲	那個Z說她因爲偷了所以就拿走了錢的皮包，就在那裡。
27	+	+	如果	那個皮包，Z說她如果偷了，就拿走錢。
28	+	–	如果	那個Z說她如果偷了就拿走錢的皮包，就在那裡。
29	–	+	然後	那個皮包，Z說她先偷了，然後拿走了錢。
30	–	–	然後	那個Z說她先偷了然後拿走了錢的皮包，就在那裡。
31	–	+	而且	那個皮包，Z說她不但偷了，而且拿走了錢。
32	–	–	而且	那個Z說她不但偷了而且拿走了錢的皮包，就在那裡。

References

Adger, David. 2003. *Core Syntax: A Minimalist Approach*. Oxford & New York: Oxford University Press.

Ambridge, Ben, and Adele E. Goldberg. 2008. The island status of clausal complements: evidence in favor of an information structure explanation. *Cognitive Linguistics* 19.3:357-389.

Aoun, Joseph, and Y.-H. Audrey Li. 2003. *Essays on the Representational and Derivational Nature of Grammar: The Diversity of Wh-Constructions*. Cambridge: MIT Press.

Aoun, Joseph, and Y.-H. Audrey Li. 2008. Ellipsis and missing objects. *Foundational Issues in Linguistic Theory: Essays in Honor of Jean-Roger Vergnaud*, ed. by Robert Freidin, Carlos P. Otero & Maria Luisa Zubizarreta, 251-273. Cambridge: MIT Press.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge & New York: Cambridge University Press.

Bader, Markus, and Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46.2:273-330.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72.1:32-68.

Bates, Douglas, Martin Maechler, and Ben Bolker. 2011. lme4: Linear mixed-effects models using S4 classes. http://CRAN.R-project.org/package=lme4

Battistella, Edwin, and Yonghui Xu. 1990. Remarks on the reflexive in Chinese. *Linguistics* 28.2:205-240.

Chen, Liang, and Ning Pan. 2003. The categorical status of finite complements of *xiangxin* 'believe' and *renwei* 'think' in Chinese. *Proceedings of the Fifteenth North American Conference on Chinese Linguistics* (*NACCL-15*), ed. by Yen-Hwei Lin, 45-53. Los Angeles: GSIL.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, Noam, and Howard Lasnik. 1977. Filters and control. *Linguistic Inquiry* 8.3: 425-504.

Clifton, Charles, Jr., Gisbert Fanselow, and Lyn Frazier. 2006. Amnestying superiority violations: processing multiple questions. *Linguistic Inquiry* 37.1:51-68.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgements*. Thousand Oaks: SAGE.

Cowart, Wayne. 2012. Doing experimental syntax: bridging the gap between syntactic questions and well-designed questionnaires. *In Search of Grammar: Experimental* *and Corpus-based Studies*, ed. by James Myers, 67-96. Taipei: Institute of Linguistics, Academia Sinica.

Culicover, Peter W., and Ray Jackendoff. 1997. Semantic subordination despite syntactic coordination. *Linguistic Inquiry* 28.2:195-217.

Dąbrowska, Ewa. 2010. Naive v. expert intuitions: an empirical study of acceptability judgments. *The Linguistic Review* 27.1:1-23.

Dalrymple, Mary, and Irina Nikolaeva. 2006. Syntax of natural and accidental coordination: evidence from agreement. *Language* 82.4:824-849.

Featherston, Sam. 2005a. *That*-trace in German. *Lingua* 115.9:1277-1302.

Featherston, Sam. 2005b. Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua* 115.11:1525-1550.

Featherston, Sam. 2007. Data in generative grammar: the stick and the carrot. *Theoretical Linguistics* 33.3:269-318.

Goodall, Grant. 2011. Syntactic satiation and the inversion effect in English and Spanish wh-questions. *Syntax* 14.1:29-47.

Gries, Stefan Th. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York: Routledge.

Grosu, Alexander. 1973. On the nonunitary nature of the coordinate structure constraint. *Linguistic Inquiry* 4.1:88-92.

Hiramatsu, Kazuko. 2000. *Accessing Linguistic Competence: Evidence from Children*'*s and Adults*' *Acceptability Judgements*. Storrs: University of Connecticut dissertation.

Ho, Man-Ying. 1995. *Interpretation of Chinese Reflexive Ziji*. Hong Kong: City University of Hong Kong MA thesis.

Huang, C.-T. James. 1982. *Logical Relations in Chinese and the Theory of Grammar*. Cambridge: MIT dissertation.

Huang, C.-T. James. 1987. Remarks on empty categories in Chinese. *Linguistic Inquiry* 18.2:321-337.

Johnson, Keith. 2008. *Quantitative Methods in Linguistics*. Malden: Blackwell.

Kayne, Richard. 1983. Connectedness. *Linguistic Inquiry* 14.2:223-249.

Ko, Yuguang. 2007. *Grammaticality and Parsability in Mandarin Syntactic Judgment Experiments*. Chiayi: National Chung Cheng University MA thesis.

Lakoff, George. 1986. Frame semantic control of the coordinate structure constraint. *Chicago Linguistic Society* (*CLS*) 22.2:152-167. Chicago: Chicago Linguistic Society.

Lee, Thomas Hun-tak. 1986. *Studies on Quantification in Chinese*. Los Angeles: University of California dissertation.

Li, Y.-H. Audrey. 1998. Argument determiner phrases and number phrases. *Linguistic Inquiry* 29.4:693-702.

Luka, Barbara J., and Lawrence W. Barsalou. 2005. Structural facilitation: mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language* 52.3:436-459.

Mendoza-Denton, Norma, Jennifer Hay, and Stefanie Jannedy. 2003. Probabilistic sociolinguistics: beyond variable rules. *Probabilistic Linguistics*, ed. by Rens Bod, Jennifer Hay & Stefanie Jannedy, 97-138. Cambridge: MIT Press.

Myers, James. 2007. MiniJudge: software for small-scale experimental syntax. *Computational Linguistics and Chinese Language Processing* 12.2:175-194.

Myers, James. 2009a. Syntactic judgment experiments. *Language and Linguistics Compass* 3.1:406-423.

Myers, James. 2009b. The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119.3:425-444.

Nunes, Jairo, and Juan Uriagereka. 2000. Cyclicity and extraction domains. *Syntax* 3.1: 20-43.

Ou, Tzu-Shan. 2006. *Suo Relative Clauses in Mandarin Chinese*. Chiayi: National Chung Cheng University MA thesis.

Phillips, Colin, and Howard Lasnik. 2003. Linguistics and empirical evidence: reply to Edelman and Christiansen. *Trends in Cognitive Sciences* 7.2:61-62.

R Development Core Team. 2011. R: a language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org

Ross, John Robert. 1967. *Constraints on Variables in Syntax.* Cambridge: MIT dissertation.

Schütze, Carson T. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Schütze, Carson T. 2011. Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science* 2.2:206-221.

Shi, Dingxu. 1994. The nature of Chinese wh-questions. *Natural Language and Linguistic Theory* 12.2:301-333.

Snyder, William. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31.3:575-582.

Soh, Hooi Ling. 2005. Wh-in-situ in Mandarin Chinese. *Linguistic Inquiry* 36.1:143-155.

Sprouse, Jon. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1:123-134.

Sprouse, Jon. 2009. Revisiting satiation: evidence for an equalization response strategy. *Linguistic Inquiry* 40.2:329-341.

Sprouse, Jon. 2011. A test of the cognitive assumptions of magnitude estimation: commutativity does not hold for acceptability judgments. *Language* 87.2:274-288.

Sprouse, Jon, and Diogo Almeida. 2011. Power in acceptability judgment experiments and the reliability of data in syntax. Manuscript. Irvine: University of California; East Lansing: Michigan State University.

Sprouse, Jon, and Diogo Almeida. (to appear). Assessing the reliability of textbook data in syntax: Adger's *Core Syntax*. *Journal of Linguistics*.

Sprouse, Jon, Carson T. Schütze, and Diogo Almeida. 2011. Assessing the reliability of journal data in syntax: *Linguistic Inquiry* 2001-2010. Manuscript. Irvine: University of California; Los Angeles: University of California; East Lansing: Michigan State University.

Stepanov, Arthur. 2007. The end of CED? Minimalism and extraction domains. *Syntax* 10.1:80-126.

Tang, Ting-Chi. 1984. *Hanyu Cifa Jufa Lunji* [*Essays on Chinese Morphology and Syntax*]. Taipei: Student Books. (In Chinese)

Wälchli, Bernhard. 2005. *Co-compounds and Natural Coordination*. Oxford & New York: Oxford University Press.

Weskott, Thomas, and Gisbert Fanselow. 2009. Scaling issues in the measurement of linguistic acceptability. *The Fruits of Empirical Linguistics*, Vol. 1: *Process*, ed. by Sam Featherston & Susanne Winkler, 229-245. Berlin & New York: Mouton de Gruyter.

Weskott, Thomas, and Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87.2:249-273.

Xu, Liejiong. 1990. Remarks on LF movement in Chinese questions. *Linguistics* 28.2: 355-382.

Xu, Liejiong. 1996. Construction and destruction of theories by data: a case study. *Chicago Linguistic Society* (*CLS*) 32:107-118. Chicago: Chicago Linguistic Society.

Xu, Liejiong, and D. Terence Langendoen. 1985. Topic structures in Chinese. *Language* 61.1:1-27.

Zhang, Niina Ning. 2009. *Coordination in Syntax*. Cambridge & New York: Cambridge University Press.

[Received 9 May 2011; revised 11 October 2011; accepted 12 October 2011]

Institute of Linguistics
National Chung Cheng University
168 University Road
Minhsiung, Chiayi 621, Taiwan
Lngmyers@ccu.edu.tw

對中文附加語及並列語孤島制約的測試

麥　傑

國立中正大學

越來越多的語法學家在例句判斷上除了考慮自身語感，也開始用實驗方式蒐集並分析無相關背景之受試者對語句的接受度。本文即以實驗測試以下三項假說：(一) 從並列語孤島中提取成分較之從附加語孤島中提取成分更容易接受。(二) 從附加語孤島中提取成分不僅在語言處理上有困難，而且也違反語法。(三) 中文的關係化由位移生成，主題化則否。本文的實驗結果僅支持這三項假說中的第一項。此外，本文所使用的語句接受度實驗成效快且操作簡便。

關鍵詞：中文，句法，孤島制約，接受度實驗，方法論