Thinking About What We Are Asking Speakers to Do

Carson T. Schütze

1 Introduction

Evidence from an increasing range of domains and types of tasks is becoming relevant and necessary to linguistic theory. Equally important is the need to evaluate the status and quality of these various types of linguistic evidence. In this paper I explore such an evaluation with respect to what we ask people to do who provide us with data – the speakers of the languages, be they consultants or participants in experiments. Specifically, I ask whether we are asking them to do things that they can understand and are capable of doing, and whether we can be confident that they are actually doing what we have asked of them. In cases where the answer to either of these questions might be No, I invite the field to seek better methods that would solidify our empirical base.

Below I examine a few case studies in considerable detail because I think that we can learn quite a lot from them. However, I want to stress that my purpose is not to criticize any particular researchers or projects, but rather to exemplify the ways in which I suggest we should **all** be continually turning a critical eye on our own research to see whether it makes as much sense as it could and should. Once one has a particular hypothesis or theory at stake, it is all too easy to persuade oneself that people can supply evidence to support it, if only they are properly cajoled. My intent is to be constructive, by highlighting aspects of experimental design that we all need to pay greater attention to.

As preliminary motivation, let us consider some classic studies on grammaticality judgements that suggest that naïve speakers, left without proper guidance, may stray far indeed from what we intend to be asking them to do. Maclay and Sleator (1960) asked beginning rhetoric students, "Do these words form a grammatical English sentence?" Three out of 21 said Yes to the string *Label break to calmed about and*. By comparison, only four said Yes to *Not if I have anything to do with it*. The authors conclude that "very little can be assumed in advance about responses to language." Hill's (1961) "request to the informants was to reject any sentences

which were ungrammatical, and to accept those which were grammatical." Concerning the sentence *I never heard a green horse smoke a dozen oranges*, two of 10 people who initially rejected this sentence "changed their votes when it was pointed out that the sentence was strictly true." Carden (1976) discovered that one of his participants was rejecting all imperative sentences because he thought that *please* should be added to them. Clearly, these studies did not have sufficient control over participants' understanding of their assigned task.[1]

2 Morphological complexity

In this section I examine some experiments on morphological complexity by Jennifer Hay. Hay pursues two claims in this area: 1) that an affixed form will tend to become semantically opaque (or "drift"), not in proportion to its absolute frequency (as is often assumed), but rather according to whether it is relatively more frequent than its base; 2) that, in positing morpheme boundaries, people are sensitive to phonotactic junctural constraints, hence these will also predict a word's potential for further affixation, in a language like English where affix stacking is highly constrained.

2.1 Decomposability of words

Hay's (2001) instructions for her experiment were highly detailed, and it is worth reproducing them here in full. Periodically I have interspersed my comments.

> This is an experiment about complex words. A complex word is a word which can be broken down into smaller, **meaningful**, units. In English, for example, the word *writer* can be broken down into two units: *write* and *-er*. *-er* is a unit which occurs at the end of many English words. In *writer*, *-er* has been added to the word *write* to make a new, more complex word *writer*. We call a word which has been made out of smaller units in this way a complex word. *Rewrite* is another example of a complex word in English. It can be broken down into *re-* and *write*.
>
> Words which are not complex are called simple words. Here are some examples of simple words in English: *yellow*, *sing*, *table*. It is impossible to break down the word *table* into smaller units. *Table* is not complex.
>
> In this experiment, you will be presented with pairs of complex words, and asked to decide which one you think is MORE complex.

In this last sentence a change seems to have occurred in how the notion of complexity is being defined: in the first two paragraphs it seemed like a categorical notion, but now we are told that there can be degrees of complexity.

> For example *happiness* is very complex – it can be easily broken down into *happy* and *-ness.*

Are we meant to contrast "very complex" here with the aforementioned characterization of *writer* as merely "complex" tout simple? I.e. is *happiness* more complex than *writer*?

> *Business*, however, is not quite so complex. While it is possible to break *business* down into *busy* and *-ness*, it does not seem completely natural to do so. *Business* is complex, but not as complex as *happiness*.

Clearly *business* is being situated between *happiness* and *yellow/table* on the scale of complexity, but there is also the claim that it can indeed be broken down. The goal for a participant at this point should be to extract what is common among the words that can be broken down that distinguishes them from those that cannot. But have the instructions provided a basis for doing so? Two potential criteria have been raised. First, in the examples *writer* and *happiness*, the complete phonology of the stem is present in the derived word, but this is not true if one regards *busy* as the stem of *business.*[2] Second, the original definition of complexity referred to decomposition into meaningful units, but while the word *busy* is meaningful, its meaning does not seem to be part of the meaning of *business*. Is the *busi* portion of *business* a meaningful chunk? I do not know what we can assume about how participants would answer this. Thus, the rules of the game seem to have changed: apparently we must now deduce some **other** justification for breaking words down, which applies when the previous two fail, if we are to match the judgements provided in the instructions. I do not know what range of possible justifications participants might entertain, but one candidate might be the fact that *-ness* is a phonological chunk of *business* just as it is of *happiness*, though its semantic contribution to the former is not so clear.

> Another example of a complex word is *dishorn*. Even though you may never have heard the word *dishorn* before, you can understand its meaning, because it can be broken down into *dis-* and *horn. Discard* is also complex – it can be broken down into *dis-* and *card*. But *discard* does not seem as complex as *dishorn*. We do not need to break *discard* into its parts in order to understand its meaning, and, in fact, it seems slightly unnatural to do so.

Here a new criterion for complexity appears to be introduced, namely the need or the possibility of understanding the meaning of the word by breaking it down into pieces. This seems prima facie different from the previous examples of complex words because they were familiar – presumably we did not **need** to (consciously) break down *writer* into *write* + *-er* in order to understand its meaning. Only upon encountering an unfamiliar word does this need arise. Now consider *discard*: what justifies the statement that it is complex? As with *busy* and *business*, the word *card* does not seem to be part of the meaning of *discard*, though *dis-* at least might be. And the instructions themselves state that it would be "unnatural" to derive the meaning of *discard* by decomposing it; in fact it strikes me as completely impossible. So, the paragraph asserts that *discard* is (somewhat) complex while acknowledging that it does not fit the criterion of complexity illustrated by *dishorn*.

> For each pair of words below, please read both words silently to yourself, and then circle the word you think is more complex. It is very important that you provide an answer for every pair, even if you are not certain of your answer. Just follow your intuition, and provide your best guess.

Apparently participants are supposed to have an "intuition" of what complexity is by this point. Since one can scarcely imagine them remembering all of these instructions, let alone inducing a consistent notion from them, the only hope for this to be true would require that participants already had the relevant concept and were simply being induced to attach the name "complexity" to it. I leave it to the reader to assess the plausibility of this assumption.

A standard response to the concern that we do not know what participants might be doing in some task is often to say that if it were not tapping something real then we would not get **any** systematic findings, just random noise. Hay does in fact get statistically significant results (by the Sign Test) in the direction she predicted (cf. hypothesis 1) above), but we can still ask how far from random the data really are. Some examples of critical items were the following (where the first member of each pair is predicted to be less complex, because the derived form is more frequent than the purported stem): *hapless~topless*, *respiration~adoration*, *rueful~woeful*, *uncanny~uncommon*, *uncouth~unkind*, *revamp~retool*.[3] Filler trials were meant to provide a baseline measure of correct understanding of the task: they paired "pseudoaffixed" and affixed words, e.g. *defend~dethrone*, *indignant~inexact*, *family~busily, adjective~protective*.[4]

Hay's results were as follows. First, her criterion for understanding the task only required responding to 20 out of the 30 fillers consistently; even with this low standard, four out of the original 20 participants had to be ex-

cluded. Furthermore, of the remaining 16, she reports that "two interpreted 'complex' in the opposite manner from that intended. This could be seen from their consistent behavior on the filler items (i.e. they rated *family* more complex than *busily*, *adjective* more complex than *protective* and so on). This consistent behavior indicates that their confusion was a terminological one, rather than a conceptual one, and so their data was included, with their answers reversed."[5] In other words, only 14 of 20 participants strictly met the rather lax standard of 67% correct responses on the filler items (chance would be 50%). This suggests there was indeed a great deal of noise in their responses, as would be expected if the instructions could not be consistently or meaningfully interpreted.

As for the critical trials, 65–66% of responses favoured the member of the pair in which the stem was more frequent than the derived form (the predicted outcome), 34–35% the opposite. Again, this should be compared to the 50% level expected by chance. Further clouding the picture, no quantitative measures of variance are reported. In particular, we should like to know whether there are individual participants who are systematically wrong on the critical items, despite passing the 2/3 threshold on the fillers, and/or whether particular items systematically contravene the prediction (in fact, some are anecdotally mentioned). Certainly there is little support in these results for the claim that the instructions in this experiment consistently led participants to any systematic basis for responding. Whether there is a way to fix the instructions depends on whether one believes people have an intuitive notion of morphological decomposability, or can learn it in short order.

2.2 Possible words

In another experiment, Hay (2002) sought to get at complexity via a different kind of intuition. The stimuli consisted of existing English words ending in *ment* to which *-al* was added, forming a nonexistent word. Participants were then to judge which of a pair of such forms is more like a possible English word, with the hypothesis being that adding *-al* is better to the extent that *ment* is not decomposed as a separate morpheme in the stem, since transparent uses of *-ment* with free Latinate stems disallow *-al* (**employmental, *nourishmental* vs. *departmental, judgemental*). Again, scrutiny of the instructions is enlightening:

> This is an experiment about possible words. You will be presented with pairs of words, neither of which are actual words of English. Your task is to decide which sounds more like it could be a word of English. Read the two

> words silently to yourself, and then circle the word you think is more likely to enter the English language.

Here we find a recurrence of the same problem noted in the previous subsection: the last two sentences of the instructions seem to be introducing different bases on which the participant is to respond. The first is based on phonology (sound) and appears to invoke a hypothetical scenario that involves the current lexicon of English. (Of course, if the word "sounds" is not emphasized, readers may well forget it and assume the instruction said "…which is more like a word of English.") The second calls for a prediction about words entering the language in the future, and does not mention phonology. If readers are no longer thinking about specifically sound-oriented properties and are approaching the general question of which word is more likely to enter the language, they could easily invoke factors such as which word expresses a more useful, frequently needed, or interesting meaning.

Perhaps unsurprisingly, the results of the experiment barely differed from chance: 56% of responses were in the predicted direction, namely that the "more likely" candidate of the pair should be the one in which the *-ment* form was more frequent than the base without *-ment* (hence, by Hay's theory, less complex). For example, *investmental* was found more likely than *arrangemental*, and *impeachmental* more likely than *enchantmental.*

3 Regular and irregular inflection

3.1 The Wug test

Berko (1958) introduced the experimental method for eliciting inflected forms of novel verbs and nouns that has come to be known as the Wug test, exemplified in (1).

(1) a. *This is a wug. Now there are two of them. There are two ___.*
b. *I often gorp with my family. Yesterday I ___ with my brother.*

When we ask speakers to inflect nonce forms, there is good reason to think that their idea of what the nonce form is supposed to represent could have a large impact on responses. For example, consider the use of the alleged default nominal plural suffix *-s* in German (Köpcke 1988; Bybee 1995; Marcus et al. 1995; Hahn and Nakisa 2000). A rough summary of its productivity in Wug tests with disambiguating sentential contexts is as follows: surnames take *-s* almost exclusively, but first names take either *-s* or *-(e)n*; nonexistent borrowings mostly take *-s* but occasionally take a zero plural; nonce common nouns and product names rarely take *-s*. Further-

more, actual recent loans (if masculine or neuter) take *-s*, but as they become integrated into the language they generally switch to a different plural inflection. Thus, in order to interpret results from Wug tests it will be crucial to establish what scenario speakers have in mind for the hypothetical existence of the novel form. Instructions that fail to make this clear invite chaos.

I suggest that this is true not just with regard to the kind of noun/verb the form is intended to represent, but also with respect to the circumstances under which the hypothetical inflected form could be part of the speaker's language. Putting the concern another way, while virtually all studies, regardless of their stand in the "Dual Mechanism debates," assume that the results of Wug tests reflect a speaker's current representation of the inflectional system of their language, I argue that this assumption must be questioned. When we ask speakers to imagine the existence of a nonce verb such as *spling*, we could be asking them two fundamentally different kinds of questions.

One invokes what I call the dictionary scenario: We are saying to the speaker, "I bet you didn't know this, but if you look in a big dictionary you'll find that English has an obscure verb *spling*; now, tell me what you think its past tense is." The second invokes what I call the neologism scenario: We are saying to the speaker, "I've just decided to make up a new verb *spling* to describe this really cool way of bouncing a paper clip that I discovered. Now, how do you think I should say it in the past tense?" (This is pretty much the scenario explicitly created for children in acquisition experiments.) It is hard to find any adult Wug studies whose instructions even hint as to which of these scenarios (or conceivably some alternative) are intended by the investigators (Haber 1976 is an exception): the instructions tend to be much more concerned with the mechanics of how the participants are to respond. Yet this is a distinction that ought to have serious ramifications for how people approach the task.

Under the dictionary scenario, it would be reasonable for a speaker to make a judgement about the shape of the extant lexicon of English, reasoning as follows: For speakers of, say, 50 years ago who had *spling* in their vocabulary along with (almost) all the current words of English that the participant already knows, how would those speakers most likely have inflected it? This is presumably what a dictionary would indicate. Analogy would likely play a large role in arriving at an answer.

Under the neologism scenario, by contrast, there are two other ways that speakers could proceed. The first would be to assume that they are in a genuine everyday neologism situation – the nonce word has indeed been coined, perhaps by the advertiser of a new product – and invoke their

knowledge of how neologisms are inflected in the language. For example, as noted above, in German most novel nouns are pluralized with *-s* when they first appear. (Whether the inflection of neologisms necessarily follows the global default of the language or whether it may be tied to a special form invoked just in "unusual" circumstances is irrelevant here.) In other words, they follow any special procedures the grammar may have for neologisms.

The second strategy would be essentially to simulate normal (implicit) vocabulary acquisition: the word is added (presumably temporarily) to the speakers' mental lexicon in the same way it would have been if they had encountered it for the first time at, say, age four. For all they would know at this age, it is just another word of the language among the many they have yet to learn, i.e. they assume it is already established in the speech community. (This is likely what happens when the second generation acquires a novel German noun and it moves from the *-s* plural class and assimilates to one of the "more native" classes.) Thus, the speaker's grammar would apply to the nonce form once the acquisition procedure had determined its lexical entry. Henceforth I refer to the two subcases just described as the neologism assimilation and neologism acquisition scenarios, respectively.

Many questions now arise, but the general point should already be clear. One would not expect a speaker to give the same response for a particular nonce word in all three of the above scenarios. If we cannot be sure which one the speaker is entertaining then we do not know what Wug test data mean. Particularly problematically, we do not know if variation within and across experiments is due to variation in the underlying grammatical systems we wish to study or in the task the speakers are carrying out.

I can do no more than speculate here on the factors that might influence which of the possible scenarios participants apply in any particular experiment. The make-up of the stimulus set is one obvious candidate. For example, Bybee and Moder (1983) mixed 16 real verbs in with 93 nonce verbs in their elicitation task; the inclusion of real words could favour the dictionary scenario for obvious reasons. Likewise, Hahn and Nakisa (2000) mixed in low-frequency real German common nouns with nonce nouns, various kinds of proper names, acronyms, etc. in their stimuli.

But perhaps the most suggestive evidence for tasks affecting the perceived scenario involves the difference between elicitation and rating tasks. In English (at least), irregular past tenses are rarely volunteered in the traditional elicitation version of the Wug test (e.g., given *spling* speakers most often produce *splinged*). However, if asked to judge the goodness of *splung* as a past tense of *spling*, they may rate it about as highly as *splinged*. The clearest data on this point are found in Albright and Hayes 2003. Their first

experiment elicited past tense forms of nonce verbs but did not ask for any ratings of those past forms. Their second experiment asked for ratings of both the regular and one or two irregular past tense forms of the same set of nonce verbs (following another elicitation for the same verb, which I do not discuss because, as the authors note, it was clearly influenced by the interspersed ratings). In the elicitation, the rate of irregular responses was just 8.7%, but the mean rating of the irregulars was 4.22 on a scale of 1 ("completely bizarre") to 7 ("completely normal"), compared to 5.75 for regulars. That is, although regulars outnumbered irregulars in production by a factor of 10 to 1, they rated less than 40% better (assuming a ratio response scale). Thanks to the authors' very detailed presentation of results, we can also look at this effect at the level of individual verbs. There were 17 nonce verbs (out of a total of 58) whose (potential) irregular past tense was never produced but was nonetheless rated at least as high as the overall mean for irregulars; these included *drice~droce, flidge~fludge, chake~chook, kive~kave*. Three irregulars that were never produced (*fleep~flept, nold~neld, chind~chound*) were rated higher than their regular alternatives.[6] One interpretation of this striking dissociation[7] could be conceived in terms of production versus comprehension: "My grammar generates *splinged*, but I would not be surprised if someone else's generated *splung*," speakers might subconsciously say to themselves, or, "I would have no trouble interpreting *splung* as the past of *spling*, if I knew there was a verb *spling* to begin with." This would result from treating the rating task as a dictionary scenario but the elicitation task as a neologism acquisition scenario.

In order to know which scenario we **want** participants to have in mind, we need an idea of what aspects of mental representations and what computations are involved in each. The neologism assimilation scenario seems the clearest: it would tell us what speakers know about how neologisms are treated in their language, but nothing more, and hence is of limited use. The dictionary scenario, I will argue, involves a frequency-weighted estimate of similarity to the existing vocabulary using a family resemblance structure. Only the neologism acquisition scenario, I claim, actually involves the speaker's grammar. I therefore suggest that the ideal Wug experiment, if it can be done, would involve guiding participants to carry out the neologism acquisition scenario. In the rest of this subsection I provide some justification for these claims about the latter two scenarios.

I maintain that the dictionary scenario is in key aspects the analogue in language of the following task from cognitive psychology. I describe to you a supposed newly-discovered species of bird found on a previously unexplored island: this bird has a large belly and a long neck, and stands three feet high. Now I ask you to guess whether or not this bird can fly (cf.

Nisbett et al. 1983). Evidently people respond in this task by applying some metric of similarity (to bird species they know about) along dimensions that they believe to be relevant to the flight behaviour of birds. Likewise in guessing whether *spling* is regular or irregular, people assess its similarity to existing verbs using properties that seem, based on past experience, to be relevant to the behaviour of verbs. The grammar may appear to be at play in this task, but this is an epiphenomenal influence: if the grammar cares about certain phonological properties, e.g. the rhyme of the final syllable, then this will be reflected (imperfectly) in the distribution of existing forms, so a general-purpose statistical mechanism can pick up on those properties as long as it has the relevant information available. Continuing the analogy, there is presumably some genuine lawful relationship between various aspects of a bird's physiology and whether or not it can fly, but the potential success of nonexperts in the flight-judgement task would not lead us to conclude that they know those physiological principles; rather, they know some superficial (partial) correlates of them. I.e., it so happens that most bird species extant today that are three feet tall or more cannot fly, but there is no physiological causal connection between this externally observable trait and flight potential; the correlation is an accident of history. (Teratorns could stand over four feet tall but could still fly.) Nonetheless, a domain-general pattern-learning algorithm can do no more or less than extract the (probabilistic) generalizations from its input. The input that feeds the dictionary scenario is speakers' cumulative experience with their language, i.e. occurrences of particular verbs in particular forms with particular meanings.

If this much is true, would we want to construct a theory of grammar based on data from a dictionary scenario version of a Wug test? Opinions differ widely on this point. Albright and Hayes (2002, 2003), for example, seem to take behaviour in Wug tasks as directly indicative of the grammar they are interested in modelling, since for them the system of rules that makes predictions about Wug behaviour plays no necessary role in producing inflected words that a speaker already knows, though those forms were the basis for the probabilistic statements that now constitute the grammar.[8] My own view is rather different: the grammar is what linguistics traditionally defines it as, namely, the mental representation of the generative system for the language we have actually acquired, and that is the object of study. The fact that we may also be able to apply very general statistical inferences, in combination with some knowledge of linguistic structure, to the generated (partial) output of that grammar based on prototypes and family resemblances, just as we can do with our observations of birds or anything else, does not imply that those statistical patterns are **all**

we represent about language, nor that those patterns even play any causal role in the grammar.

Someone who adopts my position is of course still responsible for facts that suggest that the sort of gradient similarity-based knowledge at work in the dictionary scenario also plays a role in events that clearly involve the grammar proper. Specifically, historical change and language acquisition show such influences: it is surely true, for example, that the reason irregulars tend to cluster in phonologically similar groups is that this makes them easier for children to learn, hence less likely to be regularized over the course of linguistic history. I believe this is because the language acquisition algorithm (henceforth LAA) must sometimes make guesses about the inflectional properties of words that are being acquired, and in so doing it makes use of the domain-general, probabilistic, frequency-sensitive pattern-matching abilities that I have invoked above, but that the grammar itself has no such abilities. That is, the LAA is crucially an algorithm above and beyond the "static" representation of the grammar at any particular point in time; the LAA and the grammar have different functions to perform and are subject to different demand characteristics, and they work differently. Let me sketch how this system would work in a little more detail.

To make the problem as challenging as possible, I assume the kind of grammar that Dual Mechanists and Connectionists agree cannot account for the known facts in this domain, namely a traditional rule-based system with lists of stems ("exceptions") associated with the non-default rules, à la Halle and Mohanan (1985).[9] This grammar will contain the familiar unconditioned rule saying "PAST → *ed*" and a set of more specific rules that can take precedence over it by virtue of the Elsewhere Principle, for example "ɪ → æ / __ [+nasal]…+PAST in stem class 14," which says that in the environment of a past tense suffix as the next morpheme, a high front lax unrounded vowel becomes low if it is followed by a nasal,[10] **and** this rule applies only to a set of stems identified by the lexical diacritic "class 14." This is intended to capture *ring* → *rang, drink* → *drank*, *swim* → *swam,* etc., where the stems *ring*, *drink* and *swim* will be listed as members of class 14 (cf. Yang 2002).

What happens when a child who has acquired at least the aforementioned two rules encounters a new verb and has to create a lexical entry for it? In particular, if the verb is *spling*, will it be acquired as regular or irregular (until the child hears it consistently used in the past tense, in which case it may need to change to match the input language)? In this kind of grammar, the question reduces to whether the LAA should place *spling* in class 14 or not. If it does then the grammar will unequivocally generate *splang* as its past tense. If it does not, the grammar will unequivocally gen-

erate *splinged* as its past tense; what it means for *–ed* to be the default form of past inflection is that it will be applied to stems that are added to the lexicon and assigned no additional properties by the LAA.[11] In general, how can the LAA make the decision of whether to put the new verb in any of the listed classes, and if so, which one?[12] First of all, the stem must match the phonology specified in a rule before it is even a candidate for membership in that rule's stem class. In the present case, *spling* does indeed contain the relevant vowel followed by a nasal, so class 14 is a live option; in general it might match more than one class. To decide which class to assign it to, or whether to leave it with no class specification, the LAA compares the new stem to the stems belonging to each of the candidate rule classes; in the current simplified example, the LAA looks at all the known verbs that are marked as being in class 14. The task for the LAA is to decide where the greatest similarity is, and whether the degree of similarity passes some threshold below which the new verb would receive no class specification, falling into the default paradigm.[13] I propose that in this step of the LAA the probabilistic family resemblance calculations can be employed, but not in the grammar proper. That is, without having heard this new verb in the past tense, the LAA has to make a guess about how it should be inflected.

This brings us back to the neologism acquisition interpretation of the Wug test. In fact we can now see that there are two alternative sub-interpretations possible. One is that the task involves directly adding the nonce form to the lexicon with no class marking and letting the grammar (in my specific narrow sense) apply to it. The other is that the task involves running the LAA on the nonce form, giving it whatever class status the LAA decides, and then running it through the grammar. Only the former is a pure reflection of the grammar as I have defined things. The results we expect from a task following this scenario are that nonce forms always fall into their phonologically defined default class (cf. note 11). This is pretty close to true for the elicitation version of the Wug test, as noted above: irregular responses are rare even for rhymes of irregulars. On the latter scenario where the LAA is involved as well as the grammar, we expect to see results similar to what the dictionary scenario would predict. The predicted responses are not identical, however: LAA + grammar predicts that the phonological requirements on the lexically restricted rules must be categorically obeyed, i.e. the structural description of the rule must be exactly matched, whereas a purely statistical similarity-based associative system need not require this, depending on how forms cluster in the input. I see it as an open question whether it is possible to manipulate the Wug task such that the LAA will be invoked; a possible reason it might not be is that the

LAA might cease to operate in most people after the critical period (to be replaced by a more explicit learning strategy, perhaps), so it would be unavailable to adult participants in principle. In that case, the rating version of the Wug test must be invoking the dictionary scenario, given how its responses differ from the elicitation version. If instead the LAA **can** be invoked in a Wug rating situation, this would amount to asking speakers, "What are the chances that you would acquire *spling* as belonging to the class where its past tense would be *splang*?" In neither case, however, would the rating task be a reflection of just the grammar in the way I conceive of it.

To conclude this discussion, let us step back to a more general picture. Albright and Hayes assume that the acquisition system is fairly trivial, in the sense that it does not attempt to determine which inflectional pattern a new stem should follow; rather, it simply remembers that stem, and when the time comes to produce an inflected form that has not previously been heard, the intricate probability distribution that constitutes the grammar (from their perspective) will determine what form is produced. In contrast, on the view I am following, the grammar is nothing more than was stated above, viz. rules that apply categorically given matching phonological and lexical requirements. The work of deciding how a new word should pattern is the task of the language acquisition algorithm, whose job it is to assess whether a new word matches one or more of the structural descriptions of the nondefault rules, and if so, to decide whether or not to provisionally assign that word to a class listed as undergoing those rules.[14] Their view assumes that the acquisition process is relatively simple and the grammar relatively complex, while my view assumes the opposite.

3.2 Factoring out phonology

A further methodological problem with the Wug paradigm shows itself in a study by Prasada and Pinker (1993). These researchers sought evidence that the default inflectional rule is insensitive to the phonology of the stem it is applying to, while the other inflections do show such sensitivity; this is meant to be another of the distinguishing properties of the two mechanisms in the Dual Mechanism model. In particular, the default rule should be content to apply to odd-sounding stems that do not resemble existing forms of the language, while the nondefault affixes, having no similar forms in associative memory to rely on, should resist affixation to such stems. The trick in demonstrating this, of course, is that the stems will continue to sound weird when they have been affixed to, so the target measure cannot simply be the overall goodness of the inflected form. Prasada and Pinker

tried to address this as follows: they asked their participants to rate "the likelihood of the word as the past tense of the given verb. For example, most people would probably feel that the word *drunt* is not likely to be the past-tense version of the verb *to drobe*. It is only the likelihood that the word is the past-tense form we want you to judge." Thus, they intended for participants to subtract out stem oddness in rating the derived form, e.g. judging *ploamphed* as the past tense of *ploamph* or *smamp* as the past tense of *smimp*. However, the raw data did not look as Prasada and Pinker expected: both the regular (default) and the irregular inflected forms declined in goodness as their stems got phonologically stranger.

The authors took this to mean that participants had simply failed to follow the instructions and had rated the past tense forms unto themselves, not relative to the stems. Therefore in the response data they subtracted stem ratings (which presumably reflected only the goodness of the stem's phonology) from past tense ratings, and treated the difference as their measure of interest. In so doing, they found the expected contrast: there was no additional effect of inflection on regulars, while there was on irregulars. But was their subtraction justified? Work by other researchers suggests that regular inflections **do** show sensitivity to phonological properties of stems, just like irregulars (Albright 2002), so it is possible that Prasada and Pinker subtracted away the very effect they had claimed would not occur. The interesting point in the present context is that this is the reverse of the situation mostly discussed heretofore: in this case the researchers believed speakers could **not** perform the assigned task and adjusted the data accordingly, but with less by way of evidence of inadequacy on the part of the participants than other studies where it was assumed that they **could** perform the task and data was not adjusted. Neither scenario is desirable: we want to first show independently whether people can do a task or not, rather than trying to discern this post hoc from the very data that are meant to test our hypotheses.

3.3 Rating existing inflected forms

Although rating nonce forms as possible words may be an underspecified task, it does make sense conceptually, but what could it mean to rate real words? Ullman (1999) asked speakers to rate the goodness of existing verbs in present and (regular and irregular) past tense forms in English. It should be noted that he was actually not interested in trying to interpret these ratings per se, but rather in seeing whether they would correlate with other measures in ways thought to differentiate regulars from irregulars. Nonetheless, one might wonder how much to make of data from a task for

which we have no theory of what people might be doing. Ullman does not give the full set of instructions his participants received, but here is his summary:

> Subjects were asked to give judgements based on the naturalness of the past-tense form printed in italics in each sentence. The instructions stressed that the experiments were not asking for judgements about the real-world plausibility of the sentences, but rather about the naturalness of the past-tense form in the sentence: "Is the verb in a form that 'sounds' right to you and that you would naturally use in your own speech?"… it was stressed that "it is important to remember that we are looking for your intuitions and gut feelings, and not what you believe the correct form to be according to what the dictionary says or what your teachers have told you." (p. 54)

Ratings were given on a scale from 1 ("worst") to 7 ("best"). Let us see if we can identify any potential factors on which speakers might have based their ratings. Despite the caution in the instructions, properties relating to the meaning of the sentence as a whole or to the choice of verb given the rest of the sentence may well have played a role. This could have been controlled for by having a separate group of participants rate those factors explicitly and partialing out those ratings. But focusing on what the instructions **did** ask for, there are two things in play – does the word sound right and would the speakers use it themselves? Particularly the second question invites ratings that reflect dialectal/idiolectal variation. That is, the past tense form offered by the experimenter (there was always only one for a given stem) might not be participants' preferred past form of that verb, even if they recognize it as a form commonly used by others. Indeed, if one looks at the forms with the lowest mean ratings (below 6.3 out of 7) among the irregulars this seems quite plausible. The past tense forms are *sprang, sank, shrank, wrung, bore* (stem *bear*)*, slung,* and *strung*; most of these have reasonably common alternates. A web search yields abundant simple past tense uses of *sprung, sunk, shrunk*,[15] and several of *wrang, beared,* and *slinged*; only *strung* does not seem to have an obvious alternative (*stringed* is virtually unattested).[16] Turning to the regular verbs, the lowest rated were *gore* and *pore* and their past tenses; only slightly better were the stems *jar* and *mar*. The latter two, whose past tense forms are rated much higher, might be affected by nonstandard orthographic variants *jarr* and *marr*, which are attested on the web (presumably backformed from the past and progressive forms, which double the final consonant) – Ullman makes no mention of having told participants to disregard spelling issues. Likewise for *pore*, by far the lowest rated form reported in the entire experiment, one can only guess without seeing the sentence context, but participants might

have thought the spelling *pour* or even *poor* was appropriate – two and sometimes all three are homophonous ([pɔɹ] or [poʊɹ]) in many dialects.

This is post hoc speculation, but these possibilities are readily testable. The general point is that one can easily imagine numerous factors playing into participants' interpretation of this task – indeed, for the present tense items, they virtually must invent some criteria because the instructions provide so little to go on. Ullman states that instructions for the present tense condition were similar to those for the past tense, but since the forms being rated were always homophonous with the stems, the situation was very different – no consideration of alternatives could come into play. Nonetheless, there might still have been instances (perhaps *gore* is one) where participants actually answered the question "Would you use this form in your own speech?" literally, that is, is the verb one that they themselves spontaneously produce, part of their active vocabulary? The actual target property, sounding phonologically weird vs. normal, does not seem prominent.

4 Scope and the truth value judgement task

Let us turn our attention now to the investigation of semantic knowledge. One area where it is notoriously difficult to collect consistent data involves the relative scope of two or more operators (quantifiers, modals, negation, etc.) within a sentence. Judgements of particular scope readings seem difficult even for linguists, and are often felt to be not worth even trying on naïve speakers. For example, can *Mary didn't wash two cars* mean *There are two cars that Mary didn't wash*? Asking someone, even a linguist, a question in just this form (*Can X mean Y?*) constitutes a far-from-ideal task for trying to find out what the first sentence (X) means: It requires the judger to process **two** sentences (X and Y) containing multiple scope-bearing elements, to compare the resulting meanings (of which there may be more than two if ambiguity is involved), and to deal with the almost complete overlap in lexical content, which must make X and Y and the meaning(s) associated with each hard to keep apart in short-term memory. (Much research has shown that memory for details of the surface form of a sentence is poor and short-lived, as compared to memory for meaning, but in this task it is critical to retain both.) This approach is probably too taxing from a processing perspective to yield good data, so I advocate replacing it whenever possible by an alternative, taking a methodological page from the book of child language acquisition research.

Children are more inclined to respond to the truth of an utterance than to its well-/ill-formedness. As a result, language acquisition researchers, particularly those interested in children's semantic knowledge, have tried to

get at this whenever possible by having children judge the truth of stimulus sentences that are constructed so as to allow us to infer the grammatical status of particular readings: most famously this has resulted in the widespread use of the truth value judgement task (TVJT) (Crain and McKee 1985; Gordon 1996). In this task, a scenario is acted out or illustrated with pictures, then a puppet utters a sentence attempting to say what happened; the child either rewards the puppet if its statement was a true description or punishes it (mildly) if it was false. I propose that we adopt the same basic approach with adults (minus the puppet), because the TVJT largely alleviates the problems identified with the *Can X mean Y?* paradigm above. Specifically, processing demands are reduced because only one sentence needs to be comprehended, not two. The metalinguistic component is reduced or eliminated because one no longer has to make an assessment of an utterance unto itself, but rather the task is one of describing the world and assessing the truth of such a description, a much more naturalistic way of using language. Furthermore, a visual representation of the scenario (for adults, presumably a drawing rather than a three-dimensional re-enactment) can be used to further reduce working memory demands, e.g. to keep track of the relationships between students and professors while assessing a sentence like *Some student does not like every professor*: as is sometimes done in beginning semantics or logic texts, one can draw a set of students, a set of professors, and arrows between members of these sets representing the liking relationships.[17]

As a nice example that goes part way towards applying the TVJT to a tricky scope judgement, I quote the following discussion from von Fintel and Iatridou (2003):[18]

> We are standing in front of an undergraduate residence at the Institute. Some lights are on and some are off. We don't know where particular students live but we know that they are all conscientious and turn their lights off when they leave. So, we clearly know that not all of the students are out (some lights are on and they wouldn't be on if the students were away). It could in fact be that all of them are home (the ones whose lights are off may already be asleep). But it is also possible that some of them are away. Since we don't know which student goes with which light, for every particular student it is compatible with our evidence that he or she has left. With this background, consider the following sentence:
>
> (6) Every student may have left.
> a. every student x (may x have left) true, $_{*}$ECP
> b. may (every student have left) false, $_{OK}$ECP
>
> Informants [sic] reliably judge (6) to be false in the scenario just sketched. This is predicted if the ECP [Epistemic Containment Principle] is operative.

> It would force (6) to be read with narrow scope for the quantifier *every student,* which gives rise to a reading that is false in our scenario. The ECP prohibits the reading where *every student* has scope over the modal, a reading that would be true in the given scenario. A raw truth-value judgement then supports our claim that there is an ECP.

The interpretation of this particular kind of example relies on the assumption that speakers explore all available readings of a (potentially) ambiguous sentence before answering as to its truth. With regard to children's behaviour in the TVJT this assumption has been questioned (Crain and Wexler 1999), so von Fintel and Iatridou's conclusion might be a bit hasty. Suppose instead that (6) were ambiguous, but the (b) reading is more prominent, and adult speakers do not go out of their way to try to find a means of making an utterance true in this task. Knowing that in syntactic ambiguity resolution the reading(s) not appropriate to the context often seem to be ignored without conscious awareness, we might worry whether the same is true here, in which case the reported judgements would show that reading (b) is available, but would not show that reading (a) is grammatically unavailable. I believe there are ways in which this concern can be addressed, however, and the TVJT can be put to effective use in this kind of situation.

5 Further examples

Let me briefly note some other examples of tasks used in assessing linguistic knowledge that I believe belie problems with respect to people's ability to perform the task that is asked of them and/or the interpretability of the results vis-à-vis the intended target of study.

5.1 The cloze test

One such example is the cloze procedure (Taylor 1953), widely used to assess proficiency levels of second language learners. A popular variant supplies the speaker with a narrative passage from which every n^{th} word has been omitted; the speaker is required to fill in the blanks, and answers are considered correct if a native speaker judge considers them grammatically and semantically appropriate. Here is an example where n=5:

> Pablo did not get up at seven o'clock, as he always does. He woke up late, at eight o'clock. He dressed quickly and came out of the house barefoot. He entered the garage _____ could not open his _____ door. Therefore, he had _____ go to the office _____ bus. But when he _____ to pay his fare _____

> the driver, he realized _____ he did not have _____ money. Because of that, _____ had to walk. When _____ finally got into the _____, his boss was offended _____ Pablo treated him impolitely.

The problem is that native speaker controls score far from perfect in this task. In a recent example (Cabrera and Zubizarreta 2004) from which the above passage was taken, they achieved a mean of 89% (SD = 5%), with a range of scores from 75% to 93%. Now, it is possible to compute a rating of second language learners' "competence" by measuring whether and by how much they differ significantly from native speakers as a group. Unfortunately, all we can really be certain that these scores reflect is an individual's ability to carry out this cloze task. Inferences from that to grammatical competence would require independent empirical support. The fact that native speakers show the range of scores they do (none of them scoring 100% on some passages) must mean that the task measures something other than, or in addition to, language competence. We can directly interpret score comparisons between natives and nonnatives only if we can safely assume that the groups are equal with respect to whatever those other things are that the task is sensitive to. In the present case this is very likely false: undergoing cloze tests may be a regular part of second language instruction, subject to explicit teaching and/or implicit learning of strategies specific to this task. Native speaker controls have probably never done this task before, and therefore lack any strategies or practice with the task. Thus, the two groups might achieve identical scores by different means, the nonnative group compensating for less linguistic knowledge with greater task skills. In this application, the task is not consistently measuring what it is intended to measure, because it is asking test takers to do something that some of them are better able to do than others, independent of their knowledge of the language.[19,20]

5.2 Phonology

The study of phonology is increasingly employing experimental techniques to complement traditional methods. In this subsection I briefly discuss two experiments that happen to both involve syllable structure. Smith (2004) examines people's behaviour on a loan word adaptation task: that is, she provides them with a word from a foreign language and they respond with how it would be pronounced if it were borrowed into their native language. This task is obviously parallel in many ways to the Wug test for morphology, and I would argue that many of the same concerns apply, though I shall not go through them. In addition and more interestingly, Smith pro-

vides direct evidence that speakers performing this task do not follow the same phonological rules for repairing syllable structure violations that are demonstrably at work within the their existing language. It could of course be that the experiment is correctly reflecting the presence of a distinct phonological subsystem with unique repair principles whose task is to adapt loan words, just as there could be a subsystem of morphology dedicated to inflecting novel words that plays no causal role for words that are already known. But in both cases the task has most frequently been applied with the intent of discovering **general** principles of the language as a whole, and here again it seems not to be doing what had been hoped. As in cases discussed above, the inconsistency might well arise because of the application of conscious strategies.

A more dramatic example is not so easily explained in this way, and suggests a more fundamental discrepancy between the instructions and what is actually being carried out. Brewer et al. (2004) conducted an experiment in which they asked native English speakers to respond to the question, "Is this word a possible word of English?" when presented with monosyllabic nonce forms whose onsets differed in complexity (C vs. CC vs. CCC) but all of which are legal in English (e.g., *basp, plasp, strasp*). They expected that people would rate the more complex onsets worse, by virtue of whatever it is that makes complex onsets rarer crosslinguistically. However, they found the opposite: CC and CCC forms were rated significantly **better** than C onset forms. The authors rule out several conceivable confounds before arriving at the conclusion that what speakers must be doing is rating "how unambiguously English" the forms are. That is, they were answering a question something like, "Supposing that this form is a word you do not know, how likely is it to be a word of English as opposed to a word of some other language?" Forms like *strasp* are impossible in a great many languages, so the chances that English is the language it came from are greater. (One way to test this proposed explanation would be to ask a new set of participants this question explicitly and see if the same pattern of judgements ensues.) What could explain this behaviour? A simple answer could be that the original instruction was ambiguous: it intended to ask "Is this a possible English word or an impossible English word?," but it was interpreted as "Is this a possible word of English or a possible word of some other language?," i.e., the wrong word was read as contrastive right off the bat (I believe the instructions were written, not spoken). Whether this is plausible depends on further details of the experiment that are not reported (e.g., Were examples given and how were they described? What did participants report during debriefing?). A second possibility, more along the lines of the discussion in section 3.3, is that participants

tried to follow the instructions as intended but came up with no criterion for doing so – all the forms were entirely possible, so they had to look for some other basis for providing a spread of rating responses. In other words, they either did not understand what they were meant to do or were unable to do it.

6 Concluding remarks

There is a general trade-off we seem to face in gathering linguistic evidence. On the one hand, we can use tasks for which we have a relatively clear understanding of what our experimental participants are supposed to do, and a relatively justified belief that they share this understanding and are indeed carrying out what is asked of them. Tasks such as naming, phoneme monitoring, and click location might fit this bill. These tasks tend to have the property that they are somewhat removed from normal language use; their interpretation is hence rather indirect. For example, the chain of reasoning connecting click location errors and syntactic constituency has a rather large number of links in it. On the other hand, we can choose tasks that prima facie address our questions of interest much more directly. For example, if we are interested in whether a word contains two morphemes we can ask speakers this flat out; we can require them to divide sentences into "natural groupings" of words; and so on. But then we seem to lose on the other dimension: these tasks generally seem susceptible to the possibility that naïve speakers will not understand what is really being asked of them (perhaps because we as linguists do not know how to explain it to them[21]), and/or will be incapable of doing what we ask. My purpose in this paper has been to draw attention to these dangers in this second kind of task. Our goal as a field, it seems to me, must be to develop experimental tasks that walk a fine line between these extremes. A general strategy for doing so that I have advocated is to stick as closely as possible to the ways in which language is actually used for everyday purposes, rather than contriving artificial unfamiliar tasks. In order to still gain relatively direct information about underlying linguistic knowledge, we need to become increasingly clever in the design of experiments, and I believe that doing so will be highly challenging, but in the end, highly rewarding.

Acknowledgements

For useful input on the work presented herein I would like to thank the participants in the International Conference on Linguistic Evidence, particularly Harald Baayen

and Tony Kroch, as well as Charles Yang, Bert Vaux, and the UCLA Psycholinguistics Lab, especially Colin Wilson. Most of all I am grateful to Adam Albright for discussions that have spanned many years now about Wug tests and "dual mechanism" debates, and for comments on an earlier draft. Standard disclaimers are implicit. Supported by a UCLA Academic Senate Grant.

Notes

1. There has, however, been some recent work that might be taken to suggest that, at least with respect to intuitions of grammaticality or well-formedness, how we describe the task to participants will not affect how they perform it. Specifically, Cowart (1997: ch. 4) conducted a well-controlled grammaticality questionnaire experiment that manipulated instructions as follows: One group was told to rate sentences in a way that we might characterize as prescriptive ("Indicate whether or not you think the sentence is a well-formed, grammatical sentence of English. Suppose this sentence were included in a term paper submitted for a 400-level English course…would you expect the professor to accept this sentence?…Use [the lowest rating] for sentences that you are sure would not be regarded as grammatical English by any appropriately trained person…"). The other group received instructions stressing reliance on their own intuitions ("Indicate your reaction to the sentence…Use [the highest rating] for sentences that seem fully normal, and understandable to you. Use [the lowest rating] for sentences that seem very odd, awkward, or difficult for you to understand…THERE ARE NO 'RIGHT' OR 'WRONG' ANSWERS. Please base your responses solely on your gut reaction, not on rules you may have learned…"). Cowart's basic conclusion was that this manipulation did not affect the pattern of results, though in fact he found a small but significant interaction with the other variables – he deemed that there was no linguistically interesting interpretation of this interaction. Even granting this last opinion, I would not want to draw the general conclusion that the definition or wording of a task is innocuous. Probably in this instance subjects had no prescriptive experience with the linguistic phenomenon in question, in which case they would have no reason to expect an English professor's reactions to differ from their own. For detailed discussion of the role of instructions in grammaticality judgement tasks see Schütze 1996.
2. Phonologically, *business* [bɪznəs] and *busyness* [bɪzinəs] are unequivocally distinct in the vast majority of dialects. We might assume that *happy/happiness* has taught participants that orthographic identity is not required.
3. In Experiment 3 of Hay 2000, the same instructions were used to test pairs such as the following, where juncture is the hypothesized cue to complexity: *bowlful~pipeful*, *lidless~hornless*, *punster~rhymester*, *expope~exnun*, *desalt~de-ice*, *remold~rescript*, *sheikdom~serfdom.*

4. Some of these forms might raise an issue not addressed in the instructions, namely how the potential presence of more than two affixes in a word should affect its relative complexity, e.g. the conceivable analyses *re-spir-ation, in-dign-ant*.
5. Even if one accepts this reasoning, it highlights a problem with the instructions that was alluded to above, namely that they were too long: although "complex(ity)" is the only concept introduced therein, participants apparently could not remember what it meant by the time they were finished reading them.
6. Adam Albright (personal communication) suggests that these findings do not indicate a qualitative difference in the mechanisms underlying the two tasks. He points out that, to the extent that one can make comparisons among irregulars in the first experiment (which is limited by a floor effect), their **relative** rates of production are consistent with the patterns found in the ratings (and production rates) in the second experiment. I would agree that this suggests some shared processing machinery between the two tasks. However, given similar relative rates within the irregulars in the context of very different absolute rates for them as a whole, I consider it an open question whether the common components represent a large or a small portion of the overall task machinery.
7. This task difference can also be seen in figures from Prasada and Pinker (1993), although the tasks were performed by different groups of subjects and the data are slightly obscure. In their rating task, among nonce irregulars that were most phonologically similar to existing English irregulars, five out of 10 had their irregular past tense forms rated higher than their regular past tense forms, averaged across subjects (e.g., *spling* → *splung* rather than *splinged*). Among nonce forms intermediate in similarity to real irregulars, two out of 10 had irregular pasts rated higher than irregulars, for a total of seven out of 20. Contrast this with the figures from the elicitation task, where only four of the same 20 nonce stems elicited at least as many irregular as regular past tense forms (summed across subjects, apparently); three of these were in the high similarity group, one in the intermediate group. This comparison actually overestimates the rate of spontaneous production of irregulars, because the figures represent the total number of irregular forms produced over two parts of the experiment. Part one asked for the subjects' first response, but part two allowed them to look back over all the stimuli and add any other forms that they thought would be likely. That is, if a subject's initial response to *preed* was *preeded*, in part two that subject could go back and add *pred* as an alternative, and this item would be counted as having one regular and one irregular response. As the authors point out, since there is by definition only one regularly inflected form of a given stem, but potentially numerous irregularly inflected forms, this counting method is biased to overrepresent irregular responses. (In appendix 1 of the paper, Prasada and Pinker report item-by-item "production probabilities" that suggest that, when relative rates of irregular versus regular productions are averaged across subjects rather

than being pooled, there are **no** items for which more subjects produced irregular than regular forms.)

8. This is not to say that Albright and Hayes's experiments encouraged their participants to respond using the dictionary scenario. Rather, I am suggesting that their theory seemingly implies that data from that scenario is suitable for model-building, because the theory allows for a separation between the rules that are supposed to predict Wug test data and word-specific knowledge of known inflected forms (which their model must store item-by-item for most irregulars, and can also store individually for regulars). There are (so far) no claims about how the rules and the stored items interact in the processing of known verbs: "This model is intended solely as a model of morphological productivity, and not as a model of how existing words are stored and produced."
9. I am not committing to the particular analysis that they propose for the English past tense, just the general architecture that they employ.
10. This rule will therefore not apply to *sit~sat* and *spit~spat*, which is probably correct because the class 14 stems form perfect participles in [ʌ] while these two verbs retain [æ]. My hypothesis is that if there are any phonological properties common to all members of a class, they will be reflected in the rule. Although this complicates the rule, it reduces the memory burden on the class listing because it greatly narrows the set of stems that could be members. Such phonological properties should also make predictions about acquisition, an issue I cannot pursue here.
11. In languages where inflectional class is highly predictable from phonological properties, e.g. a theme vowel, and the form the child has heard contains that vowel, the new lexical item may inflect in one of several ways because it will be caught by a rule that is conditioned phonologically but does not refer to particular stems. Opinions seem to differ on whether all of those allomorphs should be called defaults.
12. There is of course a seemingly much harder problem for the LAA, namely how to establish and refine the classes in the first place, starting from nothing. Here I am discussing a stage of acquisition when the inflectional classes have essentially reached their adult state.
13. I am crucially assuming that learners might establish a lexical entry specifying that a verb takes an irregular inflection without hearing direct positive evidence for this. Whether this is plausible for Modern English might be questioned. The fact that formerly regular verbs have on occasion innovated an irregular past tense (e.g., *ring, dig, dive*) might support this contention, though it is hard to rule out the possibility that adults made the innovation, perhaps consciously, e.g. in an attempt to sound more educated.
14. The importance of this function is more evident when considering a language like Italian, with multiple large non-default verbal inflectional classes that are reasonably productive (in the Wug sense); see Albright 2002. If the first form of a verb that a learner hears does not unambiguously indicate the identity of

the theme vowel, the acquisition algorithm must choose to temporarily assign it to one of three or four conjugation classes.

15. Indeed, there is a well-known American movie and television series called *Honey, I Shrunk the Kids*.
16. Ullman excluded five other verbs from the analysis because they had two irregular past tense forms; his criteria for establishing this were probably too conservative, however: he required both pasts to be attested in either his own dialect or one of two corpora of written text, one of which dates to the early 1960s, and the other of which is in highly uniform semi-formal style. He also states that "doublet" verbs, which allow either regular or irregular past forms, were excluded, but these were defined by relatively high ratings for the regular variant in the experiment itself, and attestation in one of the two corpora. This was probably again too stringent a requirement.
17. Another more naturalistic method being applied increasingly in the study of semantics (as well as other domains of language) is the visual-world eye-tracking paradigm (Tanenhaus et al. 1995).
18. It does not go all the way towards adopting the TVJT method because the description of the scenario in fact includes a paraphrase of one of the potential readings of the target sentence ("for every particular…has left"). Furthermore, this paraphrase has the problem that it is worded in quasi-logical terminology that would sound quite unnatural to a naïve speaker. Also, the slight oddness of "every particular student" brings to light the fact that the scenario has not fully done the work needed to make the (6a) reading potentially available: this DP wants to quantify over a set of students each of whom is known to us, but such a set is not well-established by the mere presence of the phrase "particular students" in the third sentence of the passage (thus, "**any** particular student" sounds better in the paraphrase of (6a)). These problems could be fixed by making clear that we know all the students in this residence by name (because it is the linguistics dorm, say), even introducing some names. Then instead of the paraphrase of (6a) the passage could read, "…Jenny might be out and Susie might be home, or vice versa, or Jenny and Susie might both be home and Phil might be out, etc." This does the same work of making explicit the fact that the whereabouts of any particular individual are not known, but does not require employing logic-speak or quantifiers.
19. L2 researchers might respond by saying that while this criticism may be germane to comparisons of cloze scores between native and nonnative speakers, it is not relevant to the more common application of these scores in comparing different levels of L2 proficiency. But as soon as one grants that the task is subject to strategic factors, the same argument will apply: if the groups of L2 speakers being compared have had different amounts of L2 instruction, they have probably had different amounts of training and practice with the cloze task; likewise for different **methods** of L2 instruction. The only way to escape this objection would be if there is some ceiling on the extent to which strategies and experience can improve one's cloze score, and the groups being compared have all exceeded that level of practice.

20. Sciarone and Schoorl (1989) observe that familiarity with the subject matter of the text can severely impact the results, and that tests with only 50 blanks are highly unreliable. Using 100 blanks spaced 8 words apart, native Dutch speakers achieved group mean scores of 93–97% (SD = 2.3%). Thus it may be that, **with all the parameters set appropriately**, the cloze test can identify native speakers as having (near) perfect mastery of the language. But this must be shown on a case by case basis each time a new cloze test is used.
21. Which may in turn be because we do not understand it well enough ourselves, as in the case of what precisely constitutes a morpheme.

References

Albright, Adam
2002 Islands of reliability for regular morphology: Evidence from Italian. *Language,* 78: 684–709.

Albright, Adam and Bruce Hayes
2002 Modelling English past tense intuitions with minimal generalization. In Michael Maxwell, (ed.), *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology, Philadelphia, July 2002*, pp. 58–69. Association for Computational Linguistics, New Brunswick, NJ.
2003 Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition,* 90: 119–161.

Berko, Jean
1958 The child's learning of English morphology. *Word,* 14: 150–177.

Brewer, Jordan, Lynnika Butler, Benjamin V. Tucker, and Michael Hammond
2004 *SIMPLE? Paper presented at the South Western Optimality Theory Workshop, University of California, Santa Cruz, May.

Bybee, Joan
1995 Regular morphology and the lexicon. *Language and Cognitive Processes,* 10: 425–455.

Bybee, Joan L. and Carol Lynn Moder
1983 Morphological classes as natural categories. *Language,* 59: 251–270.

Cabrera, Mónica and María Luisa Zubizarreta
2004 The role of the L1 in the overgeneralization of causatives in L2 English and L2 Spanish. In Julie Auger, J. Clancy Clements, and Barbara Vance, (eds.), *Contemporary Approaches to Romance Linguistics: Selected Papers from the 33rd Linguistic Symposium on Romance Languages (LSRL), Bloomington, Indiana, April 2003*, pp. 45–64. John Benjamins, Amsterdam.

Carden, Guy
1976 *English Quantifiers: Logical Structure and Linguistic Variation*. Corrected edition. Academic Press, New York.
Clahsen, Harald
1997 The representation of participles in the German mental lexicon: Evidence for the dual-mechanism model. In Geert Booij and Jaap van Marle, (eds.), *Yearbook of Morphology 1996*, pp. 73–95. Foris, Dordrecht.
Cowart, Wayne
1997 *Experimental Syntax: Applying Objective Methods to Sentence Judgments.* Sage, Thousand Oaks, CA.
Crain, Stephen and Cecile McKee
1985 Acquisition of structural restrictions on anaphora. In Stephen Berman, Jae-Woong Choe, and Joyce McDonough, (eds.), *Proceedings of NELS 16,* pp. 94–110.
Crain, Stephen and Kenneth Wexler
1999 Methodology in the study of language acquisition: A modular approach. In William C. Ritchie and Tej K. Bhatia, (eds.), *Handbook of Child Language Acquisition*, pp. 387–425. Academic Press, San Diego.
von Fintel, Kai and Sabine Iatridou
2003 Epistemic containment. *Linguistic Inquiry,* 34: 173–198.
Gordon, Peter
1996 The truth-value judgment task. In Dana McDaniel, Cecile McKee, and Helen Smith Cairns, (eds.), *Methods for Assessing Children's Syntax*, pp. 211–231. MIT Press, Cambridge, MA.
Haber, Lyn R.
1976 Sounding nice: Variation in English plural formation. *Kritikon Litterarum,* 5: 209–222.
Hahn, Ulrike and Ramin Charles Nakisa
2000 German inflection: Single route or dual route? *Cognitive Psychology,* 41: 313–360.
Halle, Morris and K. P. Mohanan
1985 Segmental phonology of modern English. *Linguistic Inquiry,* 16: 57–115.
Hay, Jennifer
2000 Causes and consequences of word structure. Ph.D. dissertation, Northwestern University.
2001 Lexical frequency in morphology: Is everything relative? *Linguistics,* 39: 1041–1070.
2002 From speech perception to morphology: Affix ordering revisited. *Language,* 78: 527–555.

Hill, Archibald A.
1961 Grammaticality. *Word,* 17: 1–10.
Köpcke, Klaus-Michael
1988 Schemas in German plural formation. *Lingua,* 74: 303–335.
Maclay, Howard and Mary D. Sleator
1960 Responses to language: Judgments of grammaticalness. *International Journal of American Linguistics,* 26: 275–282.
Marcus, Gary F., Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker
1995 German inflection: The exception that proves the rule. *Cognitive Psychology,* 29: 189–256.
Nisbett, Richard E., David H. Krantz, Christopher Jepson, and Ziva Kunda
1983 The use of statistical heuristics in everyday inductive reasoning. *Psychological Review,* 90: 339–363.
Prasada, Sandeep and Steven Pinker
1993 Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes,* 8: 1–56.
Schütze, Carson T.
1996 *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology.* University of Chicago Press, Chicago.
Sciarone, A. G. and Jeannette J. Schoorl
1989 The cloze test: Or why small isn't always beautiful. *Language Learning,* 39: 415–438.
Smith, Jennifer L.
2004 On loanword adaptation as evidence for preferred repair strategies. Poster presented at the International Conference on Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives, Tübingen, Germany, January.
Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy
1995 Integration of visual and linguistic information in spoken language comprehension. *Science,* 268: 1632–1634.
Taylor, Wilson L.
1953 Cloze procedure: A new tool for measuring reading ability. *Journalism Quarterly,* 30: 415–433.
Ullman, Michael T.
1999 Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. *Language and Cognitive Processes,* 14: 47–67.

Yang, Charles D.
2002 *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford.