

How we know we don't know

Shu-hao Shih (National Taiwan University)

Paul de Lacy (Rutgers University and the University of Auckland)

A colloquium talk given at the Linguistics Department at the University of Massachusetts Amherst.

3/25/2022

1. The Theory

[OPENING CRAWL]

SH: Hey, Paul.

Paul: What's up?

SH: What's a Qualifying Paper?

Paul: It's an article you have to write about some of your original research. You actually have to write two of them before you can write your doctoral dissertation.

SH: OK. I need a topic. I noticed that a lot of grad students freak out for months about finding a topic instead of just meeting with their professors and asking them about it. So, what's something good to work on?

Paul: Good question. You like doing experiments, right?

SH: Yes.

Paul: Well, bad luck. We had a great experiment guy, but he just left. On the other hand, how hard can it be? Yeah, let's do an experiment.

SH: Great. What on?

Paul: I've been working on this thing called 'sonority-driven stress'. For 16 years. Man, I'm old. Anyway, there are word stress systems where stress seeks out a vowel with a particular sonority. Kenstowicz (1993) really kicked this off. Here's really the best case I know: Gujarati.

[SLIDE: Gujarati word stress]

The most important thing in Gujarati is to stress the low vowel [a]. If the word doesn't have an [a], then you stress the second-to-last syllable.

Most stress systems just do the second part: they stress a syllable in a particular *position*. In sonority-driven stress systems, though, stress either seeks out or avoids a particular vowel. In Gujarati, it seeks out the low vowel [a]. It also avoids a schwa, but let's not complicate things.

SH: OK, but what's the theoretical interest here. I can see a *typological interest*, but what does this tell us about human cognition?

Paul: I'm glad you asked. Sonority-driven stress is *the* major source of evidence for what's called markedness conflation. It tells us profound stuff about how phonological computation actually works. On top of that, there's a flip side to sonority-driven stress. You see, a constraint like *stress/x can be satisfied by moving *stress*, or *changing x*. When you 'change x' – that's vowel reduction. So, in OT you can't have sonority-driven stress without vowel reduction and vice versa. This is called 'symmetry of effect'. You can buy my book for more details {ch-ching}.

[SLIDE: MARKEDNESS BOOK]

SH: OK, so I'm guessing that all of the descriptions of Gujarati stress have been impressionistic, right? They're reports of what people hear when they listen to Gujarati speakers.

Paul: Right. And you know, they can be a bit vague some times. People will say "Well, stressed syllables are *longer* and have higher pitch than unstressed ones", but they can't tell you *how much* longer, or how much *higher* the pitch is. So, why don't you go measure all that stuff and figure all that out. I mean, virtually no-one's done an experiment on a sonority-driven stress system before, so you're bound to get it published.

Shu-hao: Interesting.

Paul: You know, I don't want to give you the impression that I'm against impressionistic descriptions. In fact, the entire field is *built* on impressionistic descriptions. The *vast* majority of phonological theories *rely* on impressionistic descriptions, and that's particularly true of stress. I mean, this huge book by Hayes here – it's almost all impressionistic descriptions. All of my work relies on impressionistic descriptions. I mean, some people say "Oh, they're unreliable". But you know, how unreliable can they be? They've got to be hearing something, right? I mean, of course they're *reliable*.

[Sound of horse galloping]

Shu-hao: Hey, I just went to the library. Now I'm back. Were you talking to yourself?

Paul: Huh?

Shu-hao: So I got all of the descriptions of Gujarati stress together. You know, there are a *lot*.

Paul: There's more than one?

Shu-hao: And they're all impressionistic. And they all disagree.

[SLIDE GUJARATI DISAGREEMENTS]

A bunch of authors say that [a] attracts stress. But two say that it doesn't.

Also, everyone disagrees about the rest. Adenwala says that stress avoids high vowels for mid vowels. But no-one else says that. Some authors say that stress avoids schwa, but not all of them.

And by the way, they disagree about other things, like whether stress falls on the first syllable, or the penult, and whether CVC is heavy or not.

Paul: Ah, right...

Shu-hao: So when you said that impressionistic descriptions were reliable, what did you mean? Isn't 'reliable' a technical term? I thought it meant "If you repeat a method over and over again, you get the same result." But the impressionistic method is being repeated here, and we get different results.

Paul: Well, look, I'm sure Gujarati is an anomaly.

[Sound of horse galloping]

Shu-hao: Hey, I'm back from the library. So I looked at 23 sonority-driven stress languages that have multiple descriptions.

[SLIDE: Impressionistic reliability?]

They're all impressionistic. 11 of them have only one description, so we can't say anything about them. 12 of them have more than one description – they all have a "Y" in the "Replicated" column. Of these 12, 9 of them have significant disagreements in their descriptions – they all have "Y-X" in the replicated column. In other words, impressionistic descriptions of stress have about a 25% chance of agreement. At

least for sonority-driven stress systems. 75% of the time they disagree. That's really poor reliability, technically speaking.

Paul: Well, but that's *sonority-driven* stress. I'm sure there for stress *in general* the rates are a lot better.

SH: I think the real problem is that there's no way of knowing. Well, Gordon (1999) has 392 stress systems. 71% of them have only one source. In the University of Delaware Stress Database, 88% of languages have only one reference. So the lack of multiple descriptions creates an illusion of agreement.

Paul: So... let's do an experiment. I command you to go off and plan. I've gotta go do some administrative tasks.

2. The Plan: Native speakers

SH: Hey, Paul.

Paul: What's up?

SH: I need some native speakers of Gujarati.

Paul: Good news: there are tens of millions of them. And there are a bunch just north of us, in Edison.

SH: But how do we know they're actually native speakers? I mean, that their L1 is Gujarati?

Paul: Well, ummm good question. if I was writing an article and getting my data from a dictionary or grammar, I'd generally just trust that the author figured all that out. Look, I've written a lot about Maori, and based all of my work off this dictionary, basically. It's really famous. And, you know, I've never actually looked to see where the author got his data from. I mean, they don't usually tell you. {Flips it open}.

[SLIDE: HW Williams Maori sources...]

Damn. None of these are actually native speakers of Maori.

Wait! Here it is...

[SLIDE: William's Preface...]

Shu-hao: What?!

Paul: Well, look, maybe I was just being lazy. In fact, I just got one of my students to write a thesis on this stuff – Ariel Fremed. She looked at all the papers in the journal *Phonology* from 1999-2012. And, she asked “If an author cited a secondary source – a grammar or dictionary – did they report whether the source was a native speaker? The answer is 99.6%.

SH: So 99.6% of authors actually confirmed that their source reported data from a native speaker?

Paul: No, 99.6% of authors *didn't* report it.

Paul: Let's not worry about that. Let's do what we do in fieldwork instead. I've done a bunch of fieldwork. The first thing is that the person who answers your advertisement, or who the powers-that-be recommend is *not* the person you want to record. Like Dixon said:

[SLIDE: Dixon on Subjects]

- “It became clear that those Aborigines who were recommended to us by the white people were likely to be the least knowledgeable, and least useful for our purposes. The best informants tried to have as little as possible to do with whites, and were often branded by them as troublemakers.” (Dixon 1984:34)

After you find the person who wants nothing to do with you, you get to know them in their own linguistic environment. You make connections with their social group, and then through informal discussions you eventually find out if they're a native speaker. You see, some people lie about their native speaker status – for a million different reasons – political, social, financial. Sometimes they'll tell you one thing, then after they get to know you they confide in what's really going on. So, how many months will you take to get to know these speakers?

Shu-hao: Well, I'm going to put out a flyer and get them in the lab. So, I guess I'll get to know them in about 5...

Paul: 5 months should do it.

Shu-hao: ... minutes.

Paul: What the hell?

Shu-hao: Well, I can ask them questions like "Are you a native speaker of Gujarati?"

Paul: There's no way that will work. They don't know what 'native speaker' means – that's a technical term. That's jargon.

Shu-hao: Well, I *think* other experimenters ask them questions like "What language did you speak growing up?" and things like that. I don't really know because not many people actually say *how* they figured out that their speakers were native speakers.

Paul: Do they know the questions actually work? I mean, I've asked questions like that when I do fieldwork – they're the first questions you ask, and then a few weeks later, or months, you find out the opposite. I've made a list of the connection between asking someone if they're a native speaker and their responses for Māori:

[SLIDE: Responses to questions about L1ness]

You know, there were these guys who worked on Mapudungun back in the 1960s. They got 5 native speakers – what they thought were native speakers. Then after recording them they played each of the speakers everyone else's recordings. Four of the speakers agreed that the other speaker wasn't speaking their language. So, asking someone if they're a native speaker had about a 80% success rate for those guys. So, you've got to be careful.

SH: So, we either go with the 'Don't ask' method that people use when they get evidence from dictionaries/grammars, or we go with the months-long getting to know you method from fieldwork, or we go for the 'here are questions that I think will work' method.

Shu-hao: Wait a minute. What does everyone else do? How many people actually report native speaker status when they report fieldwork or experimental results?

Paul: Let's see, here we go, from Fremed's thesis again ... Ah hah.

[SLIDE: Phonology: Native speaker id/Primary]

In *Phonology*, if an author elicited data through fieldwork or experiment, they mentioned native speaker status about 90% of the time.

SH: And how many gave reasons for why they thought they had a native speaker?

Paul: 10%. So, 90% didn't give any reasons for why they thought they had a native speaker. Hmm. Well, even if you'd get away with not asking about native speaker status, you should ask. OK. I have to go do some mindless administrative task.

3. The Plan: The object of study (Vernacular)

SH: Hey, Paul.

Paul: Whassup?

SH: So I've got this human subjects protocol and I have to get the subject to sign some forms. Should the form be in Gujarati?

Paul: Well, that gets complicated. It has to go through legal, and stuff. Can't you get people who speak both Gujarati *and* English? Then the form can be in English.

SH: Yeah. In fact, everyone I've found speaks both Gujarati and English. They actually grew up speaking both.

Paul: Awesome. But, that raises a problem. You want to elicit from a *dominant* L1, not a *non-dominant* L1. A lot of people are bilingual – they have two L1s. But the L1s are *probably* always different. One's dominant, and one's non-dominant. There are real cognitive differences between them, things like reaction time, lexical access, and so on.

I've never seen anyone really say this so bluntly, but our theories are really about *dominant* L1s. Whatever makes a non-dominant L1 non-dominant is due to *extraneous performance factors* that we should factor those out. So, we gotta stick with dominant L1s.

SH: So how do I make sure that Gujarati is their dominant L1?

Paul: Well, it's really popular in fieldwork to get bilingual speakers. It makes life so much easier, especially for elicitation. But... the problem is that I've never read in fieldwork reports how they figured out whether their speaker was dominant in the target language or the elicitation language. I mean, you could maybe do lexical access speed tests. Ummm ...

Paul: And another thing: atrophy. You've got to make sure that they speak Gujarati *regularly* to *other* L1 speakers of Gujarati. If they don't, all kinds of bad things happen.

There's this case that Dixon talks about for Mbabaram where the last speaker hadn't spoken the language in decades, and each time Dixon visited him the speaker gave him different words. I guess your lexical access really gets messed up.

Apart from asking people how often they speak Gujarati and who to, I don't know of a quick foolproof method for determining atrophy, though.

SH: I'm also worried about impairments.

4. Impairments

Paul: Right. We need to make the speaker's not impaired. You know, if you think about it, there are a million ways someone could be impaired. I guess they boil down to brain impairments, articulator impairments, and hearing impairments, and then for each of those temporary vs. permanent impairments.

SH: Right. You know, some authors mention that their subjects aren't impaired, but they don't say how they figured that out, and what they used to figure it out.

Paul: Well, if you get your data from grammars, honestly, you don't worry about impairments. I guess that's not really right because there are a lot of grammars that are effectively 'rescue' operations – they get the last speaker of a language, who is often very old and full of all kinds of impairments. And that usually profoundly affects their articulation and judgements. I mean, Dunn reports that for Chukchi the old speakers couldn't do judgement tasks, for example.

On the other hand, in fieldwork, it's something I worry about a lot. But I guess I get to know what a person's like after spending a lot of time with them. I can see whether they smoke a lot, or drink a lot, and so on.

And when you look at data that comes from fieldwork, well, everyone worries about impairments. Well, what did Fremed find when she looked in *Phonology*. Here we are, Fremed looked at all the primary sourced stuff – fieldwork, mostly. And here's what she found:

[SLIDE: Properties of primary database subjects]

See, most people reported how many subjects they had.

About half reported the age.

About half reported the sex.

And a little over 20% reported mentioned asking about impairments. So, you see, worrying about impairments is important to people. 20% of them.

The other funny thing is that every subfield seems to worry about different impairments. I remember talking about alcohol with a bunch of different linguists. Alcohol can really impair your articulation, and perception, but not a lot of people actually ask about it. Some linguists laughed at the idea that you should ask your subjects if they're drunk. Others nodded and said "Of course". A professor of Alcohol studies I talked to said "Of course. The undergraduates are often hung over, especially on Monday mornings."

There are a lot of valid instruments for measuring alcoholism and inebriation. There are BAC tests, SFST sobriety tests, the MAST instrument.

SH: Well, isn't everyone potentially impaired? After all, tiredness has a lot of the same physical effects as being drunk, or being really old. So, if you're tired, you're impaired.

I actually found a circadian rhythm questionnaire. It tells you when a person should be at 'their best', which means their most alert.

Paul: Does it actually work?

SH: Yes. The authors actually tested the questionnaire and found that it was very accurate. So, it will tell you within about a 5% error rate when someone is at their best. So I'll do this test for each subject and then get them to come in to the lab when they're at their most alert.

Paul: Ahh... don't you mean 'at their *least* alert'? I was working with a speaker of Nevisian Creole when he was really alert. He said the word *person* as [peRson], which was crazy because he was non-rhotic. Anyway, we kept at it and an hour later he was showing signs of tiredness. We circled back to the word, and he said "person". The [r] had disappeared! So, we kept at it. He finally got really tired. It was just before lunch, too, and so I think he was really hungry. So, how do you say "person", I asked. [posin]. This was documented elsewhere. That formal speech is [person], informal educated speech is [peson], but vernacular speech is [posin]. So, we only got the right speech out of the guy when he was tired and hungry.

Shu-hao: Yeah but if they're tired and hungry they're likely to mumble and stumble over the words, and so on.

Paul: So, which method should we use: the alert method or tired-and-hungry method?

Shu-hao: The alert one?

Paul: Well, I don't know.

Paul: OK. I've gotta go do some administrative tasks.

5. Vernacular

SH: Hey, Paul.

Paul: Yeah, but there's something more important I wanted to talk to you about. That's how to elicit the true L1. Remember that 'dominant L1' thing I mentioned? Well, *everyone* has lots of L1s. We often mistakenly think that only people who speak multiple *languages* have multiple L1s. But of course, *everybody* speaks multiple *speech styles*.

I have a formal speech style, an informal one, and one between. And the phonologies can be pretty different.

For example, for Samoan speakers, the formal speech style has [p t k], but the informal one has [p k ʔ].

In my own speech, in my formal speech I'd say [mai neim əz pɔ:l] but in my vernacular I'd say [mæ neimz pəu]. I have a very different pronominal system in my vernacular than in my formal speech, and a bunch of phonological processes in my vernacular, like l-vocalization, coronal affrication, different vowel reduction, different flapping, and so on.

So, which speech style is the *dominant* L1? Well, it's going to be the one you primarily use when growing up – your vernacular one. The one you use when you talk to your family, and friends in informal contexts. It's not the one you learn at school, obviously. It's not the educated style.

SH: So, how do I make sure I'm eliciting the vernacular?

Paul: Well, when I get data from grammars – which is for most of my work – I guess I don't really check, to be honest. But I'm sure it's fine. I mean, I know that I need data from the vernacular speech style, so surely all my sources must have known that, too. Look, here's the grammar I used for all my work on Gujarati – Cardona (1965), and he says here {dah dah dah} “This study is a reference grammar of the standard language as taught in schools and universities and used by educated speakers” (p.7). Damn.

I'm sure this is just an anomaly. I mean, look, here's another grammar – a really famous one by Skorik (1961) on Chukchi. So much has been written about Chukchi in phonology. Epenthesis, Syllabification. It's got a sonority-driven stress system, too! And what does Skorik say – you know, I've never checked before...

[SLIDE: Skorik]

.... “literary Chukchi”. But that doesn't mean that people don't speak literary Chukchi natively. Except I see here that Dunn (1999) says that it doesn't. Damn.

Shu-hao: You know, aren't dictionaries and grammars *always* about the formal speech style? Unless they explicitly say that they're about informal speech, then shouldn't we assume they're about some literary high style that no-one actually speaks?

Paul: Fine. Let's put aside grammars and dictionaries. Anyway, when I do fieldwork, I spend months – sometimes years – cultivating a relationship with my consultant. I get to know them as friends. And eventually that pays off.

But you've only got 5 minutes with these speakers. Damn.

SH: Well, maybe I could achieve it with my environment.

Paul: Brilliant idea. Make them feel at home. That's a good start. Where are you recording them by the way?

SH: Well ... I'm going to take them into a basement, then a small windowless room, then into a smaller windowless room, then put them in a steel booth and wire them up.

Paul: I'm sure that will make them relaxed.

SH: OK. I've got an idea. I've got a research assistant who speaks Gujarati. I'll get her to chat with each of the subjects for about 10 minutes. She can talk to them about their family and informal stuff like that.

Paul: Great idea. Let me know how that works out.

Some time later...

SH: Hey, Paul.

Paul: How did it go?

SH: Well. The problem was that my RA is female. And some of the subjects were male. So, I don't know if she got them relaxed, or whether they got nervous.

Paul: I guess we should have measured their heart rate.

SH: Also, after talking to them, I put them in the windowless room in the steel booth anyway, so I don't know how long the vernacular effect lasted.

Paul: Right.

SH: And some of the words I elicited might have been pretty formal.

Paul: Right.

SH: But I made my carrier sentences have vernacular words in them.

Paul: Right.

SH: But, look, I've recorded them now, so let's move on.

Paul: Fair enough. Talk soon. I've got to go to an 'important' admin meeting.

6. Measuring

SH: OK, I've got a bunch of sound files, and Praat. So, it's time to measure. Any advice on which method I should use?

Paul: Well, I guess you're interested in the vowels, right? So, from what I recall, you open up the Praat, and squint at a spectrogram. And when the spectrogram goes kind of dark, that's the beginning of the vowel.

SH: Right. No, I mean that there are five different ways of finding the beginning of a vowel after aperiodic noise. Which of those do you think I should use?

[SLIDE: Methods for identifying]

Paul: 5? Which ones are those again?

SH: Well, I could use the WAV method {demonstrate}. Or the F1 method {demonstrate}. Or the F2 method {demonstrate}. Or the F3 method {demonstrate}. Or the F0 method {demonstrate}.

Paul: Well, which one does everyone else use? I think I've heard a lot of people mention F2.

SH: Good news! I've found a paper that compares the methods. Francis et al 2002. They used an electroglottograph to figure out exactly when the larynx closed – when voicing began after a stop. They then compared all of the methods to the electroglottograph results. Here's what they found.

[Slide: Stop+vowel boundaries: Accuracy, Precision]

It really depends on whether you're measuring after an aspirated vs. unaspirated stop. ...

Look at the WAV method. It's very precise. And It's pretty accurate. It has a precision of just a few milliseconds, and it has an accuracy of only a couple of milliseconds or so.

The F2 method is good for unaspirated stops, but not so good for aspirated ones. Notice that it's pretty inaccurate and imprecise.

Paul: Hang on a second. What do these terms 'accuracy' and 'precision' mean? I mean, I throw around the word 'accuracy' a lot, but I'm guessing that they're actually being used as technical terms here.

SH: Right.

[SLIDE "Accuracy & Precision"].

Suppose you do a lot of measurements, or shoot a lot of arrows.

If all of the arrows hit around the same point, then you're really precise. If all your measurements are close together, you're really precise.

And if all of the arrows hit around where the target, you're really accurate. If all your measurements are clustered around the correct value, then you're accurate.

You want a method that's *both* precise and accurate.

Paul: You know, I think until this moment, we've been talking about problems but not coming up with solutions. The fact is, we've got a clearly defined theory of sonority-driven stress. That's set within a well defined theory of the phonological module, and even a broader well-defined theory of linguistic architecture. We've also got observations. We've got observations in grammars and dictionaries, and from impressionistic judgements from fieldwork, and from experiments.

Those aren't our problems. The things we keep on shrugging our shoulders about are how these things are connected. It's about how we get from our theory to evidence for our theory, no matter where it comes from. We have methods, but we just don't know whether those methods work. Or rather, how *precise* and *accurate* and *reliable* those methods are.

SH: Yeah. This is the first time in our discussion when we can actually say "Here's a method we'll use as part of getting evidence for our theory, and we know how accurate and precise it is. We know that it's valid.

Paul: It's the first time we haven't shrugged our shoulders. I mean, without this work by Francis et al., we wouldn't know which of these 5 methods was best, right?

SH: Right. And even the best method isn't perfect. The best method is still a couple of milliseconds away from the gold standard.

Paul: So, this is really heartening and disheartening at the same time. It's obvious that when we talk about evidence for our theory, we need to spell out our certainty about that evidence. And thanks to Francis et al we *can*. We can say "We measured using the wav method, and so we know that each measurement had a precision of a couple of milliseconds, and its accuracy was within a couple of milliseconds."

SH: Right, we're quantifying uncertainty. So, if we get measurements within 2 milliseconds of each other, we know that we shouldn't take that difference too seriously.

The funny thing is, I remember reading papers where people claimed that there was a statistically significant difference between stressed and unstressed syllables of less than 5ms, but some of these methods aren't precise enough to know that this difference is real.

Paul: OK. Awesome! So, how do you find the end of a vowel?

SH: Well, there are many different methods, but I can't find the same kind of study for that. Lots of people use F2, but I don't know how accurate and precise that is.

Paul: Oh. OK. Good luck. I've gotta go do some admin.

7. RAs

SH: Hey, Paul.

Paul: What's up?

SH: You know, measuring is taking ages. Could you lend me some research assistants to help me?

Paul: Well, sure. I have lots of RAs. But, you know, I guess we should worry about their accuracy and precision, right?

SH: Yeah. Lots of people do – inter-rater reliability.

Paul: I actually just got a bunch of RAs to go through some glossolalia recordings for me. I asked them to find all the stops. Simple, right? After I did that, I thought: I wonder how accurate and precise they were. Here's what I found.

[SLIDE – Find the stops]

SH: It looks like just using one undergraduate will get you around 75% of the stops.

Paul: Right, with a range of between 64% and 82%.

SH: So the precision is 18%.

Paul: Yeah, so if you use one undergrad, they might find as few as 64% of all the stops, or as many as 82%.

SH: So, to get all of the stops, you had to use 6 RAs?

Paul: Well, there's a *chance* you'll get all the stops if you use 5 RAs. If you want to be guaranteed to find them all, then you need 6, right.

SH: That's a lot of RAs.

Paul: Yeah. But look at the problem if you do it yourself. There's no reason to think that you'll perform better than the best RA. So if you do it on your own you'd only find about 82% of the stops. So, it's obviously better to use lots of RAs.

SH: Damn. I better get all my RAs together, then.

Paul: Wait! I did a bit more. I then got the RAs to find the beginning of the vowels, using the WAV method – because it seemed like the most accurate and precise one.

SH: Great.

Paul: I got all 6 RAs to do the measurements. Here's what happened:

[SLIDE: Precision in the WAV method..]

SH: So, this graph shows the range of inter-rater agreement per number of tokens. So, looking at the bottom edge of the graph, we can see that there were a couple of tokens where every rater agreed absolutely.

Paul: Now, if you look along the x-axis, you can see that the RAs agreed with each other within a range of 5 milliseconds or less for about 44% of the tokens.

SH: So, this tells me how much precision I can expect. For any token, there's an 82% chance that the raters will agree on where the vowel starts within 11 milliseconds or less.

Paul: Right. Cool, eh? So, another way to put it is that if you grab any word, there's a 50% chance that everyone agrees where the vowel begins, within a range of about 6 milliseconds. There's a 93% chance that they'll agree within 19ms.

SH: OK. So, doesn't that mean that we have to take precision into account when we report our results? I mean, shouldn't every statistical result be reported with a note about the precision of the method you used? Like, if someone says "Stressed vowels are on average 100ms long", well, you would *have* to say "But our method meant that for any measurement there's a 90% chance we could be 10ms off either way."

Paul: You know, that reminds me of a lot of results we've been seeing.

SH: Yeah, I've just read Matt Gordon's (2004) paper. It says that there's a 10ms difference between primary and secondary unstressed vowels. But I think only one person measured the vowels. So, how sure can we be that this 10ms isn't due to imprecision?

Paul: The difference between primary and secondary stressed vowels are even closer – like only 8ms. I don't know if these differences are real, or just a record of the method's imprecision. But most of the papers I've read are like this.

SH: So do I factor your findings into my results?

Paul: I don't know. The problem is that we might have to measure this kind of reliability and precision for each new group of RAs you use. RAs are human, and humans vary a lot, so it's not clear that my results here matter for any other group of RAs.

SH: Damn.

Paul: Yeah. OK. I better go do some administration.

8. Stats

{I really don't think we need to talk about this. Let's see how much time we have}.

9. Results

[SLIDE: Shih 2018 - *Phonology*]

Paul: So, tell me about the Gujarati results! What's stress like in a sonority-driven stress language?

SH: Your question is undefined.

Paul: What?

SH: There is no sonority-driven stress in my data. Stress always falls on the second-to-last syllable. Look..

[SLIDE: F1 & F2]

[a] always has exactly the same vowel quality, no matter where it appears.

[SLIDE: Gujarati F0]

All words always have the same pitch pattern.

[SLIDE: Gujarati Duration]

Duration doesn't match, either.

Paul: But, that completely disagrees with the impressionistic descriptions ... which all disagree with each other. But even so, all of the impressionistic descriptions – well, except for a couple – agree that there's some sonority-driven stress. So, what's going on?

SH: Well, [a] always has a long duration because of the amount of time it takes to lower the jaw. I guess that if you don't hear the kind of acoustic cues that you're used to hearing in your own language, you start paying attention to inherent distinctions. So, I think that linguists couldn't hear cues like pitch and vowel centralization, and so they seized on the fact that [a] is longer than all the other vowels.

Paul: So, you're saying that when people talk about Gujarati stress, they're actually reporting that [a] has a longer duration than the other vowels?

SH: Yes.

Paul: But Gujarati was the best case of sonority-driven stress. It had the most data, and the most descriptions, and the most number of actually living, accessible speakers. And I wrote a bunch of things about it – de Lacy 2002a and b, 2004, 2006, 2007. I think my career just flashed in front of my eyes.

But! But! There's no way that this could hold of *every* sonority-driven stress system. I mean, I've found like 25 of them. And in a lot of them, stress avoids *schwa*. I mean, *schwa* and stress *hate* each other. There *must* be systems where stress avoids *schwa*.

Look, where there is smoke, there's fire. Right? Right!

Time passes...

SH: So, I did some fieldwork, some more experiments, and I gathered up all the cases I could – like about 40 of them. And I wrote my dissertation.

None of the cases work out. I mean, some of them are just misdescribed. There are conflicting impressionistic descriptions for a bunch of them. The experimental results don't match up with the impressionistic observations.

Paul: Do you secretly hate me?

SH: It's not just me. A couple of other people have looked at sonority-driven stress independently. Bowers independently looked at Gujarati – no sonority-driven stress there, which agrees with my findings. And Raisin wrote a dissertation where he looked at a bunch of the descriptions and found that they didn't actually support the claims.

Paul: But what about the systems where stress avoids *schwa*? I mean, there have *got* to be cases like that.

SH: Well, Vartan Haghverdi has looked at Armenian, which is supposed to be like that. No actual experimental evidence there. I think people are just used to associating vowel reduction with unstressed syllables, and so they hear a *schwa* and think "That must be unstressed".

I looked at Paiwan, and found that there are actually three types of *schwa*: long, short, and supershort. Anyway, stress doesn't avoid *schwa* there either.

[SLIDE: Non-moraic schwa PhD diss.]

Paul: I thought that where there was smoke, there's fire. But apparently sometimes there's just smoke. I think I need to lie down. And do some admin.

10. Re-evaluating

SH: Hi, Paul.

Paul: So, I've been thinking about what this all means. You've destroyed my life's work. And it's not just that.

I mean, my theory was so beautiful. I had an algorithm that created constraints that combined prosodic elements at all levels of the prosodic hierarchy with sonority levels, and tone. And Chris Oakden has just destroyed tone-driven stress, by the way, so that's that, too.

And then there's the symmetry thing. I mean, if there's vowel reduction, there has to be sonority-driven stress. That just follows from basic theoretical assumptions. But if there's no sonority-driven stress, then how can there be vowel reduction? Is there vowel reduction? I don't even know anything any more. I know absolutely nothing.

SH: You have become like Socrates.

Paul: So I should quit and become a farmer? Fair enough.

I keep on thinking about what went wrong. You know, the funny thing is that I've always worried about evidence for phonological theories and how to get it. In fact, I wanted to write my doctoral dissertation on evidence for phonological theories, but ended up writing about markedness.

And I think we've come to the clear conclusion that unless we know the validity of our methods, we can't claim to know anything with any certainty. Knowing the accuracy and precision of our methods lets us say "OK, I know that I'm a few milliseconds away, or what chance I've got of this person not being a native speaker, or not speaking in their vernacular." And more than that, it tells us when we can say "OK, that's not a real difference, even if it's statistically significant, because it's within the range of precision."

I wonder why method validity hasn't been central to my research.

SH: Well, to be fair I haven't seen it discussed very often. You get a few papers that talk about it, but you have to hunt for them. And it depends on the subfield. If you look at people who rely on grammars and dictionaries for their data, they almost never talk about their methods, and they don't talk about method validity.

There are *some* though. Blumstein & Stevens (1979) talk about the accuracy of a method for determining place of articulation. Lahiri et al (1984) talk about another method that has a 91% error rate.

There are also studies on the accuracy of certain forced aligners, like the p2fa.

Paul: And they should, right? I mean, suppose I say "Māori has no trisyllabic words with more than two heavy syllables in them, and I know this because I looked in this dictionary." Well, no-one should trust that. I would have to spell out my method. *How* did I look in this dictionary? Did I scan it then OCR it, then do a regular expression search. Or did I thumb through the pages while sitting on the loo? Those are completely different methods, and they're going to have different levels of accuracy. So, what are those levels? I think if anyone says "There are *n* words of this shape", then you have to ask "What's the accuracy of your method?" I haven't found any studies of dictionary-searching methods that look at their accuracy and precision. I hope there are some. But if there are, I haven't seen anyone *factor in* the accuracy and precision of these methods in their research. And that seems crucial, right?

SH: Yes, but there's also the 'weak link' issue. The fact is that getting any kind of evidence requires applying multiple methods. It's a composition of methods. You have to figure out if the data was elicited from an unimpaired, non-atrophied, dominant L1, right? Each of these involves multiple instruments, and they'll be different if you're using a dictionary, fieldwork, or experiment as sources. The fact is that if you don't know the validity for any method in the chain of methods you use, you can't quantify the overall validity of the evidence.

Paul: So we have to know the validity of every method in order to know how much to trust our evidence.

SH: The funny thing is that it's not really that *hard* to study method validity.

Paul: It doesn't seem to be. Francis et al's paper is great, and there are a lot of papers measuring the validity of methods for impairments. In some ways, this work is half done for us; it's just about re-examining it in the context of linguistics and cognitive science, and maybe tweaking it here and there.

SH: Right. For any method, you just need a gold standard, then compare methods to that gold standard.

Paul: So, working on method validity is *doable*; it's not even that *hard*. But I wonder if people will actually do it.

I think there are cultural reasons why people would resist.

I remember talking to a syntactician – a theoretical guy. He'd done some fieldwork and was telling me about some methods he'd used. I said "Those are brilliant. Can I read about them in the paper you wrote?" He said "I didn't write about them because every word I spend talking about my method is a word I don't have to talk about my theory."

I think there are subcultures in linguistics which see methodology as something that's not essential. Something you can cut out. In phonology, it's weird: if you get your evidence from experiments, you *must* include a methods section. If you get it from fieldwork, it's optional. If you get it from grammars and dictionaries, there's never a methods section. That doesn't make any sense. These are just different evidentiary modalities. They all involve a process – methods – so all require us to talk about them.

SH: I've noticed that, too. By the way, remember that paper we wrote about the evidence for sonority-driven stress – the one we sent to the *Catalan Journal of Phonology*?

[SLIDE: Evidence for sonority-driven stress]

Paul: Yeah. You mentioned that someone sent you comments. What did they say?

SH: Here's the main objection: "Are we all supposed to become experimentalists now?"

Paul: I don't get it. What?

SH: I think the problem is that as soon as you start talking about methods, a big bunch of people just say "They're talking about experiments." It's like different rules apply depending on your evidence source.

Paul: But of course, even impressionistic judgements could be useful tools. It's just that they have to be quantified carefully. How accurate and precise, and reliable, are impressionistic judgements for reduplication, say? Or l-vocalization? They might be very accurate and precise, for all we know.

SH: I think some people think that as soon as you say "experiment", theoreticians are out of a job.

Paul: I don't get it. Theoreticians are essential. How do you get evidence for a theory without knowing what the theory is? An essential part of any method is proving that it has *construct validity*, which boils down to proving that it derives from a theory. It has to make theoretical sense. Doesn't matter how accurate or precise it is.

SH: Well, maybe if everyone started working on method validity, we'd actually get theoreticians, field workers, and experimenters working together seriously. After all, methodology is the thing that connects theory with evidence, so maybe it will make everyone collaborate more.

Paul: I think that was the thinking behind Laboratory Phonology, to be honest.

You know, I think another issue is going to be the subfield-specific attitudes about which evidence matters. You know, if you want to publish a paper on intonation, you have *got* to have done an experiment. You can't say "Impressionistically, intonation is like this..." But if you work on *tone*, impressionistic judgements are fine. I remember a tonologist telling me that she had gone to an intonational conference and reported some impressionistic judgments of intonation, and the intonationists had shouted her down. She was really angry. I remember criticizing a semanticist for talking about impressionistic intonation, but not hassling a tonologist for reporting impressionistic judgments.

Isn't that funny? We have tiny subfields in phonology that have completely different methodological standards.

SH: Well, there are big divisions even among phonologists about what counts as phonology. Do you remember that paper I submitted on Gujarati sonority-driven stress? The one that reported the experiment? The reviewers loved it. It didn't even have to go through a second round of reviews.

Paul: That was probably due to brilliant supervision.

SH: Yes. Well, I sent another paper to another journal about another language. It also presented a phonological theory and an experiment as evidence. The result: "This is phonetics".

Paul: What was phonetics?

SH: The experiment. For some people, as soon as you say "experiment" it's phonetics, not phonology.

Paul: So, you propose a phonological theory, and you say "And here's the evidence for my theory." And if the evidence is from a dictionary, or grammar, or impressionistic judgements from fieldwork, or from yourself, it's still phonology. But if the evidence is from an experiment, it's phonetics?

SH: Right.

Paul: But for those people, wouldn't that mean that almost *everything* is phonetics? If you're a syntactician and your evidence hinges on prosodic phrasing, so you measure someone's pauses, then that's phonetics, too, right? If you're a semanticist, and the evidence for your theory hinges on where an intonational pitch accent lands, so you measure pitch, then you're doing phonetics, not semantics, right?

How on Earth did we get to the stage where there are divisions between phonologists simply based on your source of evidence?

SH: OK. So we know what to do. What now?

Paul: Well, the problem is that there's only one of you. This issue is too vast for any individual to fix. We have so many methods to examine.

SH: Right, but it's necessary. We can't really know anything much without knowing the validity of our methods, right?

Paul: Right, but that's a hard sell to a grad student. If you had come to me and said "I want to write a QP on method validity", I would have said "That's a bad idea." Why? Because where are the papers on method validity? Show me them in *Phonology* or *NLLT* or *LI*. I think this reflects the value that the field places on method validity. It doesn't.

Look, I'm sure that everyone would say "Of course method validity matters". But when a good chunk of the field doesn't even *mention* the methods they use in their research, the writing's on the wall.

And I think it's been this way for a long time. Like I mentioned, when I went to write my PhD in 1999, I said to John McCarthy "I want to write about evidence for phonological theories." He said that that was a terrible idea, and I wouldn't get a job. I don't know if anything's changed.

SH: So who *can* work on method validity?

Paul: Not people who are getting tenure for the same reason. Maybe after you get tenure. But the reality is that that's when they dump admin on you until you drown. And then they hassle you about becoming a full professor, which makes you retreat into things you know – sure bets. And then you wake up and you're 50 and your whole body aches, and then they make you chair. So, maybe if you're 55 or 60.

SH: Interesting.

Paul: So, ... anyway. Where does that leave us on sonority-driven stress? We have impressionistic evidence, which we know is unreliable. And we have experimental evidence, which disagrees with it. And we know for some of the methods involved what their validity is, but not for others – maybe the majority of the methods. So, do we actually know anything to any good degree of certainty?

SH: Of course we do. We know that we don't know.

Paul: *And* we know why we don't know.

SH: A great achievement.

Paul: Yeah.

[SLIDE: Outtro]

[SLIDE: Thanks, references]

END

=====