

## **Data and Grammar: Means and Individuals**

*Marcel den Dikken, Judy B. Bernstein, Christina Tortora & Raffaella Zanuttini*

### *1. Introduction*

In the abstract of his target piece, Featherston states that ‘it is no longer tenable for syntactic theories to be constructed on the evidence of a single person’s judgements’. In our commentary, we focus on this issue and on (what we perceive to be) Featherston’s claims that individual speakers’ judgments are intrinsically unreliable and ‘noisy’, that variation among individuals’ judgments should be smoothed out by averaging the judgments of a large pool of informants, and that only this golden mean counts as genuine data. We argue that these claims are at odds with the basic premises of Chomskian linguistics, which is centered on the I-language of the individual speaker/hearer, not the E-language of the speech community. This is not to say that the way generative grammar approaches and accumulates its data is in no need of improvement: we will make some specific recommendations of our own to this end in the last section of our commentary.

### *2. Data and data gathering in generative linguistics*

In every field of scientific inquiry, it is important to have a clear definition of the object of study and the methodology used to investigate it. For generative grammarians, from the inception of the framework, the object of study has primarily been the speaker’s knowledge of his or her native language. On the way to this ambitious goal, the researcher is confronted with an immediate and obvious difficulty: How can (s)he gain access to the knowledge of a native speaker, which is, after all, subconscious? What counts as the data that can be investigated in order to build a model of the speaker’s knowledge of his/her language? Linguists within the generative community have been relying on the introspective judgments of native speakers concerning whether a given sentence is acceptable or not (what is usually called, with a misleading label, a ‘grammaticality judgment’). Depending on whether or not the speaker finds a certain sentence acceptable, the linguist concludes that the grammatical model to be built must or must not be able to generate it, and modifies the model accordingly. Because the object of investigation is the knowledge of an individual, the researcher could work with a single native speaker and characterize his/her knowledge of language, or grammar. However, since speakers live in communities that in a naive sense share a language, it is common practice to check whether the judgments provided by an individual are shared by a number of other individuals as well, and, if so, to conclude that what is being built is a model of the grammar that is ‘shared’ by the speakers of that ‘language’. Though plausible, this assumption is also problematic, and is at the root of what we perceive to be a series of misunderstandings, as we will show later in this commentary.

More than five decades of research in generative linguistics have shown that the standard generative methodology of hypothesis formation and empirical verification via judgment elicitation can lead to a veritable goldmine of linguistic discovery and explanation. In many cases it has yielded good, replicable results, ones that could not as easily have been obtained by using other data-gathering methods, such as corpus-based research. Think, for instance, of the phenomenon of parasitic gaps: it was generative grammar that led to the discovery of these creatures and that developed highly detailed insight into their distribution. Traditional linguistic data-mining, perhaps most impressively represented by Otto Jespersen’s work from the early 20th century, unearthed a wealth of data and insight. But consider the fact that parasitic gap constructions, which are exceedingly rare in corpora, are restricted to A-bar dependencies, are subject to an anti-c-command condition, and can only be licensed by S-structure movement: these distributional phenomena would have been entirely impossible to distill via any non-introspective, non-elicitation based data-gathering method. Corpus data simply cannot yield such a detailed picture of what is licit and, more crucially, what is not licit for a particular construction in a particular linguistic environment.

Despite the undeniable successes, however, it is often thought that generative linguistics has a ‘data problem’. The use of elicited judgments has been criticized by variationist sociolinguists because it cannot capture information concerning the different rates at which speakers use certain forms as opposed to others.

This is clearly a problem if different frequencies of use are a reflection of linguistic knowledge, as argued by some (for example, Emily Bender or Gregory Guy; see, e.g., Bender 2005; Guy & Boberg 1997). The criticisms raised by Featherston are of a different nature, and reflect the opinion of some linguists who feel that individuals' judgments lack objectivity and reliability and that overall judgment elicitation does not represent a sufficiently 'scientific' kind of methodology. We very much welcome the opportunity to reply to Featherston's article, as a way to engage in this important discussion. We will start by addressing some general points raised by Featherston that echo criticisms more broadly encountered within the field; then, in the next section, we will discuss what we perceive to be at the core of Featherston's thinking. Generative syntacticians are sometimes accused of not taking data seriously. Featherston says that they are 'reluctant to examine the data in any detail' (p. 1) and even show a 'disdain for evidence' (p. 8). It is worth reflecting on what gives rise to this kind of criticism, and to what extent it is warranted.<sup>1</sup> One reason for saying that generative syntacticians do not take data seriously seems to be the fact that they use introspective judgments, which are considered to be subjective (i.e., possibly reflecting the whim of an individual), and not worthy of being treated as objective pieces of data, on a par with entities that exist outside of the individual's mind. We do not share the view that something that is in an individual's mind is not real or not to be treated as a legitimate object of investigation. Generative syntacticians, generative linguists in general, and cognitive scientists even more generally all work with individuals and with subjective material, as that is what they must necessarily rely on, also in the construction of their 'objective' experiments and their selection of informants.<sup>2</sup> Does this mean that their data base is necessarily flawed? Imagine the limiting case of a linguist working on his native language and using only his own subjective judgments as 'the data'. Of course there is a risk that his theoretical predispositions and working hypotheses could color his judgments.<sup>3</sup> But unless this linguist speaks a variety that nobody else in the field speaks, he will not get away with fabricated data: material submitted for publication will be vetted by other native speakers of the variety investigated, to the extent that they are available; rigorous scholars remain aware of this, and act accordingly.

What, then, of languages for which there is at most one native-speaker linguist available? Could facts from such languages be used as evidence for a theoretical claim? Featherston says that 'isolated examples from obscure, little-studied languages' are an 'unsatisfactory data source' because '[s]uch data is uncheckable, the original data source was probably superficial', and the author 'probably knows no more of the language than exactly this point which they have taken from a descriptive grammar' (p. 7). We do not feel comfortable with a generalization that characterizes descriptions of 'little-studied languages' as

---

<sup>1</sup> It is true there are some pieces of work that make sweeping theoretical claims without sufficient emphasis on the data; but if that work is worth its salt, it will be vindicated in detailed empirical studies of the kind that there is certainly no shortage of in generative circles. The generative linguistic literature contains a wealth of careful examinations of linguistic data. It is also true that linguists have on occasion been too hasty in reading into the judgments they obtained the existence of different dialects or idiolects (see our discussion below of "Dutch A" versus "Dutch B"); but it is important to bear in mind that these are honest mistakes that are generally corrected in later research. In the discussion that follows we will not focus on what we take to be mistakes or problems caused by hasty researchers; we concentrate instead on problems that are seen as stemming from the kind of data and methodology used within generative linguistics.

<sup>2</sup> Featherston's own work provides examples of this practical need: in Featherston (2005a:688, 691) certain sentences are not tested because it is doubted that they 'will reveal any additional relevant effect' (p. 691): '[o]ur own intuitions ... reveal no ... differential effect' (p. 688); '[o]ur own judgments detect no difference' (p. 691).

<sup>3</sup> Presumably because of this concern, Featherston says that '[i]t is simply inadequate research practice for linguists to rely on their own unconfirmed introspective judgments as linguistic evidence' (p. 7).

‘superficial’. In our work on lesser-known languages, we have found that the data in descriptive work often has great depth and, more importantly, is almost consistently confirmed by linguistically naïve native speakers (see e.g. Den Dikken 2003:73, with specific reference to Rotuman). Moreover, while it is certainly true that an isolated datum from an unverified source for a language that one does not know is not an adequate basis for a theory, to our knowledge the literature contains no published work that presents such an isolated datum as the *sole* piece of support for a major theoretical claim. There exists, on the contrary, a lot of detailed work on lesser-known languages that presents large clusters of data obtained via careful informant work.

To be sure, such work does not always involve ten or more informants working their way through a paper-and-pencil or computerized questionnaire. Featherston voices the concern that generative syntacticians often gather their data through informal conversations with just a handful of speakers, and proposes that ‘ideally we should gather the judgements of twenty or thirty speakers, but as few as ten or twelve will suffice for some purposes’ (p. 11). We agree that it is desirable to work with data elicited from as large a number of informants as possible. However, the generative syntactician’s reason for desiring many (versus one) informants would differ from that given by Featherston: for the generative syntactician, the more informants you have, the more data from individual grammars you have, which gives you the potential to find micro-variants you might otherwise not have found (this is not unlike the general desire to study as many ‘languages’ as possible). However, despite the desirability of working with data elicited from a group of informants, we do not believe that the field should *require* that every piece of research be based on the judgments of at least ten or twelve speakers. For many of the world’s languages, it would be extremely difficult to recruit ten or more informants for a questionnaire study. Moreover, a paper-and-pencil or computerized questionnaire is not always appropriate. A written questionnaire of the type suggested by Featherston, though perfectly feasible for written languages with literate informants, would not work for languages that lack a written tradition or are perceived as stigmatized varieties of standard languages (see Cornips & Poletto 2005, Henry 2005, and Bernstein 2006 for relevant discussion). Similarly, the laudible goal that each construction type under examination be tested using a number of lexical variants is not always achievable, due to limitations on the part of the informants. For example, many aging informants can devote only a limited amount of time to a task, and they often represent the only group of native speakers of certain languages.

Perhaps with similar considerations in mind, Featherston does not lay down his ‘Essentials’ as hard-and-fast rules. To us, they are at best desiderata; if they cannot be met in a particular informant study, they should not be taken to be a basis on which to discard that study as a valuable data source. After all, all data are data (except, of course, for fabrications), nothing more but certainly also nothing less. Suppose that a particular judgment on a particular set of data from a lesser-known language such as, say, the Polynesian isolate Rotuman, replicates in minute detail (*modulo* irrelevant differences, of course) facts from densely-studied languages such as English or Italian. Should such data be cast aside if they were obtained from fewer than a handful of informants, or should they be presented as evidence for a particular analysis of the construction at hand and chalked up as support for the (UG-based) theory within which this analysis is couched? Past experience has shown that the former would be ill advised because the latter yields reliable and replicable results. Of course past performance is no guarantee of future results, and every case needs to be considered for its own merit; but in general, we see no reason to set aside data obtained from few informants of some little-studied language as useless.

So far we have addressed broad criticisms that have been raised against the kind of data used in generative linguistics, some of which are also found in Featherston’s target article. Now we are ready to turn to what we take to be the heart of Featherston’s paper: the explicit proposal that we should not treat as data the judgments of individuals, but rather the mean value of the judgments of a group of individuals. This proposal raises questions of a fundamental theoretical nature; we turn to these in the next section, which is the centerpiece of our commentary.

### 3. *Individuals or groups? Noise or variation?*

At the heart of Featherston's criticism of the data and methodology used in generative linguistics is the reliance on the judgments of *an individual*. The core of his proposal is that we should rely instead on the judgments of *a group*. In this section, we will examine the reasoning on these two points and offer a detailed discussion of the problems they raise.

Featherston states that, by relying on the judgments of individual speakers, linguistic research 'excludes the possibility that there is a universal grammar' (p. 8). He argues that '[t]he big questions in generative linguistics are those which refer to all speakers, not just one speaker, and to the whole language, if possible to all languages, not just to a single lexical string' (pp. 29–30). We do not agree with the view of Universal Grammar that these statements express. Universal Grammar in Chomsky's I-language sense is not the set of Greenbergian (E-language) universals, that is, the set of surface properties that are common to sets of languages. Universal Grammar is the abstract grammatical system that, by hypothesis, is innate in the mind/brain of all humans. Every speaker's grammar is, once again by hypothesis, a reflection of Universal Grammar, and, as a result, every speaker's grammar must meet the requirements imposed by the invariant principles and the particular parameter-settings allowed by UG. Given this, relying on the judgments of individual speakers certainly should not be incompatible with studying UG. Research has shown that humans (even 'savants' with a special talent for learning languages; see Smith & Tsimpli 1995) are radically unable to acquire a language that does not conform to UG's requirements, suggesting strongly that it is indeed the case that UG imposes severe limits on the grammar of even a single individual. UG lives in every individual, and there is in principle nothing wrong with trying to model the grammar of a single individual as a way to study its limits and possibilities.

We do not share the skepticism that we detect in Featherston's article toward working with a single individual. Suppose that one had in one's possession a body of published literature that covers a range of syntactic phenomena found among a large group of speakers, and that one wanted to do follow-up fieldwork on the language in question. Imagine further that one started out by interviewing just a single speaker. To verify if this speaker is 'representative of the speech community', one would ask him/her for judgments on all of the syntactic phenomena that previous published research has reported on. Suppose that this speaker in fact reports judgments that are entirely consistent with what the extant literature reports. On these data, then, this speaker's judgments are perfectly in line with the group judgment. Suppose now that one asked this speaker for judgments about sentence types that have not been investigated in the extant literature. It does not follow that those judgments would have to be mistrusted, even discarded as subjective and unreliable, simply because they were obtained from just this one speaker. In fact, the opposite follows: if all of the previous judgments of this informant were consistent with those represented in the literature from a large number of speakers, then the judgments given for data not yet discussed in the literature are to be readily trusted.

This said, there can of course be no doubt that eliciting the judgments of a number of individuals is an extremely useful way to investigate correlations among properties in a grammatical system. To this extent, all generative linguists would agree with Featherston that studying a group is useful. All would probably also recognize (again, with Featherston) that, when we generalize from a group of speakers to all speakers of that language, we are making a gross overgeneralization, since we know that grammars can vary from one individual to the next. The point at which it gets harder to agree with Featherston is in the claim that studying a group is the *only* way to do solid work. He states that '[t]he group produces a clear, statistically significant pattern revealing a syntactic generalization, but the judgements of each individual informant are noisy and much less visibly systematic' and therefore 'the real appropriate measure is between any single informant's judgements and the *mean value of a group of informants*' (p. 12; original italics). We disagree with these assertions, for the following reasons.

Generalizations over a group can only be made if the individuals that constitute the group share the same linguistic knowledge. Yet there is no a priori reason to expect each individual belonging to a certain group of speakers to have set every single parameter in exactly the same way as his/her fellow group members. Since we know that all human languages change over time, we are forced to conclude that different members of the same ‘group’ of language speakers internalize subtly different grammars. Such grammars are obviously very close, ensuring perfect mutual intelligibility, but may differ in some fundamental ways in their I-language constitutions, thus leading different speakers to analyze a particular string in different ways, with each particular hypothesis leading to different analyses of other input strings and, importantly, to different output utterances regulated by the parameters in question. This is the received view, within the generative framework, of how language change comes about (see work by David Lightfoot and Anthony Kroch; e.g., Lightfoot 1999; Kroch 2001), and we find it to be an entirely sound and sensible perspective.

Since one cannot know in advance whether one’s informants all have the same I-grammar, it is difficult to define a group over which one can confidently take an average. Informants may all be native speakers of what we naively take to be the same ‘language,’ say German, as in many of Featherston’s case studies, but that does not mean that they all share exactly the same parameter settings (significant differences in fact do exist among German speakers with respect to a host of grammatical features). Prior to running an experiment on a particular construction type, it is typically hard or even impossible to know or forecast what kinds of variation one might encounter for that particular construction type within a group of informants that all ‘speak the same language’. What one typically finds is that there is linguistically significant variation among speakers with respect to their judgments on individual sentences in the questionnaire. Averaging the informants’ responses to a mean value will obliterate individual differences. The problem with this is that all potentially interesting points of variation are then cast aside as ‘noise’, and the net result is lots of gray averages. Universal Grammar thus becomes Universal Gray, and that would hardly be reflective of the real patterns that micro-analysis would allow one to identify.

To be sure, the ‘error bars’ in Featherston’s graphs (whose length shows the 95% confidence interval for the mean) do give one a global sense of the range of variation in the judgments. But they cannot reveal individual speakers’ patterns or the systematic sub-regularities that might be hidden underneath the surface means. To appreciate the nature of the problem, consider Figure 1 in Featherston’s target article: the means for the four cases reveal a clear pattern, one which seems at odds with Grewendorf’s claim. But the error bars are extensive and they all overlap robustly, so it is entirely possible that there is a subset of speakers in the set whose judgments are averaged in Figure 1 for whom ‘NPacc<sub>j</sub> sich<sub>j</sub>’ is better than ‘NPacc<sub>j</sub> ihm<sub>j</sub>’ while ‘NPdat<sub>j</sub> sich<sub>j</sub>’ is *worse* than ‘NPdat<sub>j</sub> ihm<sub>j</sub>’ — in other words, for whom the Grewendorf judgment holds. Taken by themselves, therefore, the means in Figure 1 cannot be taken to ‘not confirm Grewendorf’s intuitions’ (p. 5).<sup>4</sup> Rather, the fact that the error bars are long and overlap extensively should be a signal to the experimenter that the means may hide grammatically significant speaker variation.

In general, we disagree with Featherston that it is *always* the case that ‘the judgements cluster around a mean’ and that this mean represents ‘the “underlying” value, free of the noise factor’ (p. 11). To us, this amounts to taking all variation among informants to constitute noise or errors. In contrast, we are convinced that grammatically significant variation among informants exists. Featherston seems to assume homogeneity within a group of informants — a kind of optimism (or, viewed from the opposite angle, pessimism about the existence of genuine dialectal and idiolectal variation) that may be the accidental result of the particular case studies he conducted. But just around the corner from the case studies he reports on in

---

<sup>4</sup> As a side point, we do not see how the mean value of the judgments of a group of speakers can confirm or disconfirm an individual’s judgments: one’s judgments are one’s judgments, no matter what other speakers of ‘the same language’ might think.

his paper (and also in his recent publications; see Featherston 2005a,b), one finds genuine and undeniable variation. Take for example the Dutch data in (1):

- (1) a. ik vraag me af [*wie* Jan gekust heeft]  
 b. ik vraag me af [*wie dat* Jan gekust heeft]  
 c. ik vraag me af [*wie of* Jan gekust heeft]  
 d. ik vraag me af [*wie of dat* Jan gekust heeft]  
 ‘I wonder who John kissed’

As shown by highly meticulous recent questionnaire studies conducted as part of the *Syntactic Atlas of the Dutch Dialects* (Barbiers *et al.* 2005) project, there really are differences among individuals when it comes to the choice between the ‘standard’ form (1a), featuring just the *wh*-word *wie* in the left periphery of the embedded question; the form with the declarative complementizer *dat* ‘that’ to the right of *wie* (1b); the form with the interrogative complementizer *of* ‘if’ following *wie* (1c); and the form which combines the interrogative and the declarative complementizer (1d). While some speakers allow several or even all of these forms, the distribution of (1b–d) is restricted in ways that one would fail to notice if one averaged over the entire Dutch-speaking world. Such variation may be dialectal (clearly geographically determined) or idiolectal — it is often hard to tell. But whatever the nature of the variation, it is clear that it is genuine variation.

It must be acknowledged that linguists have on occasion been too hasty to read into the judgments they obtained the existence of different dialects or idiolects. A classic case in point is the infamous “Dutch A” vs. “Dutch B” distinction (Maling & Zaenen 1978), based in part on one of the phenomena singled out by Featherston in his paper: the *that*-trace effect. It has become clear in more recent research that there really is no dialect split on this point within the Dutch-speaking world, and that hence the “Dutch A” vs. “Dutch B” distinction is not real (see esp. Bennis 1986:sect. 3.6.1). This was an honest mistake — and moreover, not one that was uniquely the consequence of relying on the subjective judgments of just two informants. But though there is no “Dutch A” vs. “Dutch B” distinction, it remains true nonetheless that there are differences, when it comes to the (im)possibility of *that*-trace sequences in Dutch, between individual types of sentences, depending on the presence or absence of non-verbal material between the complementizer and the verbal cluster, and (in the absence of such material) word order in the verbal cluster.<sup>5</sup> Thus, for many speakers, the *that*-trace examples in (2a,b) are better, regardless of the word order in the verbal cluster, than (2c), which does not have either an object or a locative PP preceding the verbal cluster. Example (2c) in turn is somewhat better than (2d), which features the past-participial main verb to the right of the finite auxiliary.

- (2) a. wie denk je dat *ec* de samba {gedanst heeft/heeft gedanst}?  
 who think you that the samba {danced has/has danced}  
 b. wie denk je dat *ec* in de keuken {gedanst heeft/heeft gedanst}?  
 who think you that in the kitchen danced has/has danced  
 c. wie denk je dat *ec* gedanst heeft?  
 who think you that danced has  
 d. wie denk je dat *ec* heeft gedanst?  
 who think you that has danced

<sup>5</sup> The general gist of this conclusion can be found in Bennis (1986:237–38). It is confirmed experimentally by a questionnaire study conducted by the first author of this commentary, which includes 66 example sentences tested on more than a dozen informants (all linguists). Since we do not have space in this commentary to include the details of the study and its results, we refer to the following website: <http://web.gc.cuny.edu/dept/lingu/dendikken/papers.html>.

There is nothing about the presence or absence of non-verbal material between *dat* and the verbal cluster or the order inside the verbal cluster *per se* that is marked. These things play a role only in environments in which the structural subject position of the embedded clause is unfilled. Even then, their effect is not one that can easily be couched in terms of a dichotomy between grammatical versus ungrammatical: none of the examples in (2) are radically unacceptable, but there are *relative* contrasts.

These facts on the one hand help confirm something Featherston says in his paper: that judgments on individual, isolated examples could very well skew the result of the empirical investigation, and steer it in the wrong direction. For instance, what if the *that*-trace examples with intransitive embedded verbs that we checked with our informants had all been of the (2d)-type? In isolation, many informants might very well have rejected them, because, in truth, they are less than brilliant; others, however, might have accepted them. We might have concluded from this that there are two dialects of Dutch: “Dutch A” and “Dutch B”, one resisting subject extraction across *dat*, and the other allowing it. But we would have been wrong to do so: extraction of the subject across *dat* is in fact grammatical for all speakers; but specific instances of such extraction are worse than others. In this regard, then, Featherston’s cautionary notes are entirely well taken.

There is another side to the coin as well, however. Suppose we follow Featherston’s advice, give informants a variety of different examples in a random order, and then compute the mean scores for all sentences involving *that*-trace violations. We are bound to miss what turns out to be an interesting pattern in the examples: the fact that (2d) is relatively worse than (2c), which in turn is relatively worse than (2a) and (2b), where the order in the verbal cluster is inconsequential. Unless we actually do a micro-analysis of the data, looking at individual examples and at the individual speaker judgments on these examples, we end up concluding that *that*-trace sequences in Dutch are grammatical, though, on average, never quite perfect. But that would be the wrong conclusion to draw: there are subtypes of clauses in which *that*-trace sequences are perfectly fine.

That computing the mean value of the judgments can be misleading is shown perhaps even more clearly by the Dutch comparative correlative construction (cf. English *The more you read, the less you understand*). The results of an extensive questionnaire study reported in detail in Den Dikken (to appear: Appendix) show that, in this domain, one finds a kind of variation in the judgments that would get completely snowed under in a Featherston-type mean value approach. On individual sentences, one finds speaker judgments to range from totally impossible (\*\*) to perfectly fine (OK), with just a little bit of grey (? , ?? , ???) in between. Typically, speakers are quite categorical about their judgments, and equally typically, speakers categorically disagree amongst themselves with respect to the status of individual examples.

What are we to take away from this? It seems to us that these kinds of results tell us two things. First of all, they bring home the importance of taking individual judgments seriously and considering them microscopically, rather than computing the mean value of the judgments of the group. Such averaging would lead to a profound misunderstanding of what is going on.<sup>6</sup> Secondly, they show that speakers typically are not afraid to go for one of the extremes on the rating scale provided. This is true not only for linguist informants; it is perhaps even more emphatically true, in our experience, for naïve native speakers. Most of the time, individual speakers are confident whether they can say a particular sentence or not — and they often are happy to volunteer meta-linguistic comments indicating what led them to accept or reject a particular sentence, something we will come back to in our closing remarks. It is of course true that speakers sometimes find a particular sentence so-so rather than jolly good or woeful — but a ‘so-so’ judgment is *not* a hedged judgment: informants use their assortment of question marks to indicate, usually *with confidence*,

<sup>6</sup> We disagree with Featherston’s (p. 16) reading of Chomsky (1965:3); Featherston states that for Chomsky, ‘[t]he variation of the individual’s knowledge from the norm, personal quirks in usage, and *dialect variation* are irrelevant variables which are to be controlled for’ (emphasis added). We do not think that Chomsky casts aside dialect variation as irrelevant.

that the sentence is neither perfectly fine nor perfectly bad. It is then incumbent on the analyst to derive this judgment from a theory in which multiple principles and parameters as well as levels of grammatical representation beyond ‘narrow syntax’ have jurisdiction over the fate of individual utterances.<sup>7</sup> But whatever the theory decides, linguists’ use of intermediate judgment diacritics to indicate the pre-theoretical status of the stimuli is a healthy practice: it is part and parcel of the essential process of full disclosure.

#### 4. Working with individuals

*A propos* this process of full disclosure, we certainly agree with Featherston that it is important that linguists reveal their sources, as well as their data. Having access to the entire data set may in many cases be vital, both at the review stage and beyond, for readers to be able to probe into possible causes of variation in the judgments on specific sentence types. In this way, confounds resulting from inadvertent generalizations over tokens assumed to represent the same type can be brought to light. There is an added bonus to such full disclosure as well: the test items for the experiment at hand could very well be of great value to others who are looking into a different grammatical aspect of some of the token sentences. These test items could serve as ‘raw material’ for a different experiment on an entirely different topic, perhaps in an entirely different field of inquiry, including neighboring fields of linguistic inquiry as well as other cognitive sciences, such as psychology.

In the foregoing, we have made some remarks regarding, among other things, Featherston’s requirement to use multiple informants and multiple tokens, his ‘mean value’ approach to the data, and his perspective on variation as ‘noise’, and we have recommended that authors make their full set of experimental data available in their papers. On this last point, we would actually like to go further, and add the recommendation that linguistics journals adopt the practice of requiring full disclosure of data sources in publications of empirically based work, ideally in the form of an appendix in which the full data set is revealed. In addition, we have some recommendations to offer regarding methodology and the procurement and use of data. These recommendations are prompted by experience emerging from the present authors’ ongoing research and fieldwork on the comparative morpho-syntax of Appalachian English, and also from fieldwork with speakers of non-standard Romance varieties conducted by some of the present authors. For discussion of the history behind these approaches to fieldwork on the syntax of non-standard varieties in Europe, we refer the reader to Benincà (2004) and Cornips & Poletto (2005). Thus, the recommendations we offer have behind them a rationale supported by a history of successful fieldwork on the syntax of non-standard varieties. We also refer the reader to Henry (2005), who offers many insights, some of which we discuss below.

Let us recall the fact that working within the generative framework means that we take I-grammar to be our object of study, and furthermore that we must often rely on the intuitions of native speakers to provide insight into grammatical structure. What this means is that when working with informants, it is very helpful to try to gain insight into what has moved them to accept or reject examples. To this end, it is highly advisable for the experimenter to engage in a discussion with informants. Often informants have a keen meta-linguistic sense of what is wrong with sentences they are presented with, and it can be extremely helpful for the linguist to be privy to these meta-linguistic judgments (Henry 2005). For example, informants often volunteer insightful ways of improving particular sentences, and they often point out that a particular sentence they just heard is perhaps not entirely impossible but significantly worse than one they heard previously (or *vice versa*).

---

<sup>7</sup> We do not see the logic in Featherston’s claim (p. 9) that those using a Chomskian categorical theory should use only clearly (un)acceptable examples as evidence. It is not true that ‘this model was explicitly designed to deal only with “clear cases”’: there is more to language than ‘narrow syntax’ alone.

On a related matter, it should be noted that informants can be ‘trained’ to reflect on their own grammar in the same way that linguists are trained to do so. Thus, before testing the target structures on a potential informant, it is essential to work on structures that are well-known to be possible (or impossible) and to familiarize the informant with what the linguist ultimately intends by ‘acceptable’ and ‘unacceptable’. In such a training session, the informant learns the protocol, and is asked to give judgments of stimuli that the experimenter knows are ungrammatical in the language (like ‘word salads’ of the type *Grandma the happy is*) or frequently occur in the informant’s spontaneous speech (like *Ain’t nobody happy*, a non-standard English construction that is used frequently in many American varieties). An interaction of this sort is extremely valuable, for a number of reasons. It puts in context the meaning of elicitation questions (‘How natural does this sound?’, ‘Would you use this?’, ‘Could you say this?’), downplays the importance of the particular elicitation question used, helps the informant get a sense of what participation in the experiment amounts to, and, most importantly, allows the experimenter to find out in the training session which informants are likely to be able to give usable judgments and which are simply incapable of doing so. This is an important tool that is readily at the experimenter’s disposal to help eliminate some of the real noise that could come out of a questionnaire study.

We also advocate engaging the informant in discussion, for a number of reasons. As we noted, a trained informant will often volunteer unsolicited opinions about the relative acceptability of two similar sentences. These *relative, comparative* judgments are extremely valuable to the linguist, and underscore why an interactive conversation is necessary. Discussion with the informant is also important because it allows for control of unexpected semantic or pragmatic interpretations. So, although stimuli should be presented in carefully constructed contexts to control for these unexpected interpretations, an informant may nevertheless judge a sentence as unacceptable, not because of ungrammaticality, but because the informant had a particular interpretation in mind that the researcher could not have imagined or predicted. These discussions with the informant can clarify the reasoning behind the unexpected judgment, providing insight that might not have been ascertained otherwise. It is thus important to bear in mind that a good informant can act as a collaborating linguist, offering suggestions for what led to a particular judgment.

Finally, we cannot emphasize enough the importance of taking individual judgments seriously. ‘Outliers’ should not be cast aside as ‘noise’ as a matter of course. Whenever an apparent ‘outlier’ presents itself, one should try to ascertain whether it might be correlated with some other ‘outlier’ from the same informant. What matters, after all, is what patterns there are in the data. One apparent ‘outlier’ may team up or correlate with another apparent ‘outlier’ elsewhere, and the correlation between these ‘outliers’ may be enormously revealing from a theoretical point of view. Thus, ‘outliers’ are potentially linguistically significant, and one would not want to miss the opportunity to be educated by such ‘outliers’ by making them all come out in the wash. Of course none of this is to say that there is no such thing as ‘noise’. But just as one cannot know what ‘a group’ is before one has encountered it (in the form of judgments that pattern together), one also cannot know in advance what constitutes ‘noise’. So one should treat all feedback as potentially linguistically significant, setting it aside as ‘noise’ only if it cannot fit into any kind of meaningful pattern.

### *Acknowledgment*

This material is based upon work supported by the National Science Foundation under Collaborative Grant Nos. BCS 0617197, BCS 0617210, BCS 0616573, and BCS 0617133.

## References

- Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Magda Devos, Margreet van der Ham. 2005. *Syntactic Atlas of the Dutch Dialects*, volume 1. Amsterdam: Amsterdam University Press.
- Benincà, Paola. 2004. "Dialetti d'Italia e dialetti d'Europa," *Quaderns d'Italià* 8/9: 11-26.
- Bender, Emily M. 2005. "On the boundaries of linguistic competence: Matched-guise experiments as evidence of knowledge of grammar," *Lingua* 115(11), 1579-1598.
- Bennis, Hans. 1986. *Gaps and Dummies*. Foris: Dordrecht.
- Bernstein, Judy B. 2006. "How speaker judgments inform the study of Appalachian English syntax," unpublished ms., William Paterson University.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.
- Cornips, Leonie, and Cecilia Poletto. 2005. "On standardizing syntactic elicitation techniques (part 1)," *Lingua* 115:939-957.
- Dikken, Marcel den. 2003. *The Structure of the Noun Phrase in Rotuman*. Munich: LINCOM Europa.
- Dikken, Marcel den. to appear. Comparative correlatives and successive cyclicity. To appear in Anikó Lipták (ed.), *Correlatives: Theory and Typology*. North Holland Linguistics Series. Amsterdam: Elsevier. (Ms. available on <http://web.gc.cuny.edu/dept/lingu/dendikken/papers.html>)
- Featherston, Sam. 2005a. "Universals and grammaticality: Wh-constraints in German and English," *Linguistics* 43.4, 667-711.
- Featherston, Sam. 2005b. "That-trace in German," *Lingua* 115, 1277-1302.
- Guy, Gregory & Charles Boberg. 1997. "Inherent variability and the OCP," *Language Variation and Change* 9.2: 149-164.
- Henry, Alison. 2005. "Non-standard dialects and linguistic data," *Lingua* 115: 1599-1618.
- Kroch, Anthony. 2001. "Syntactic change," in Mark Baltin and Chris Collins (eds.), *The Handbook of Contemporary Syntactic Theory*. 699-729. Malden, Mass.: Blackwell Publishers.
- Lightfoot, David. 1999. *The Development of Language: Acquisition, Change and Evolution*. Oxford: Blackwell.
- Maling, Joan and Annie Zaenen. 1978. "The nonuniversality of a surface filter," *Linguistic Inquiry* 9.3, 475-497.
- Smith, Neil and Ianthi-Maria Tsimpli. 1995. *The Mind of a Savant: Language Learning and Modularity*. Oxford: Blackwell.