EXTRAORDINARY CLAIMS REQUIRE EXTRAORDINARY EVIDENCE (AND ORDINARY ONES REQUIRE ORDINARY EVIDENCE): ON EXPERIMENTAL LINGUISTICS FOR LESS WELL STUDIED LANGUAGES

Uli SAUERLAND
Zentrum für Allgemeine Sprachwissenschaft, Berlin, Germany (ZAS/Berlin)

ABSTRACT

The late physicist Carl Sagan, whom I quote in the first part of my title, skillfully phrased the common sense view on evidence in the mature sciences. In linguistics, however, evidence has become a controversial issue, especially so when it comes to the investigation of less well studied languages. In this paper, I argue that Sagan's principle should be applied to linguistics. The growing accessibility of a wide array of experimental techniques and computational tools to analyze such data makes it feasible to back up extraordinary claims with evidence from a variety of sources. At the same time, it is in many cases possible to agree on what constitutes an ordinary claim and focus the extra effort on extraordinary claims. For non-controversial claims no more than the minimum effort to establish the claim and properly document the evidence is necessary.

RESUMO

O falecido físico Carl Sagan , citado na primeira parte do título deste artigo, formulou com habilidade a visão do senso comum sobre a natureza da evidência nas ciências maduras. Em linguística , no entanto , a evidência tornou-se um assunto controverso, especialmente quando se trata da investigação das línguas menos bem estudadas. Neste artigo, defendo que o princípio de Sagan deve ser aplicado à linguística . A acessibilidade crescente a uma

grande variedade de técnicas experimentais e ferramentas computacionais para analisar dados linguísticos torna viável apoiar propostas extraordinárias a partir de evidências de uma grande variedade de fontes. Ao mesmo tempo, é , em muitos casos, possível chegar a um acordo sobre o que constitui uma proposta científica comum, não extraordinária, deixando para concentrar qualquer esforço extraordinário apenas para apoiar propostas igualmente extraordinárias . Para propostas não controversas não é necessário mais do que um mínimo de esforço para se estabelecer e documentar as evidências.

KEYWORDS

evidence, fieldwork, syntax, semantics, methodology

PALAVRAS-CHAVE

evidências, trabalho de campo, sintaxe, semântica, metodologia

Introduction

Evidence has become a topic generating substantial discussion in theoretical and descriptive linguistics. For instance, the University of Tübingen in Germany organizes a biannual conference entitled "Linguistic Evidence" since 2004. The conference describes itself in the call for papers for the 2014 conference as "a meeting place for linguists who wish to improve the empirical adequacy of linguistic theory and linguistic analysis." and aims to "more closely integrate data-driven and theory-driven approaches"[1]

Implicit in this description is the view that the empirical adequacy of linguistic theory is open to improvement because the theory has not paid sufficient attention to the accessible linguistic evidence (data). Also several recent journal contributions focus on the methodology of collecting evidence to address questions in linguistic theory. I discuss

[1] http://www.uni-tuebingen.de/forschung/forschungsschwerpunkte/sonderforschungsbereiche/sfb-833/ev/le2014/call-for-papers.html, accessed Aug. 2, 2014

below a debate concerning evidence in theoretical syntax and semantics focusing on English data (GIBSON and FEDORENKO, 2010, 2013; SPROUSE and ALMEIDAM, 2013; SPROUSE et al. 2013), but also two recently published contributions on the methodology of fieldwork (MATTHEWSON, 2004; DIXON, 2007). I hope to show that despite all the discussion within linguistics, the same view towards evidence used in the established sciences can also be applied in linguistics. SAGAN (1980) aptly phrased this principle as I quote it in the title of this paper: "Extraordinary claims require extraordinary evidence".

My paper is structured into three sections. In the first section, I articulate some general principle relating to evidence. In particular, I show the Sagan's principle has been the common view of the relationship between theoretical claims and empirical evidence for centuries dating back at least to HUME (1748). The second part of my title is in parenthesis ("and ordinary ones require ordinary evidence") because it remains an implicature in Sagan's formulation. But I show that it has been understood as part of the principle since the beginning. I then provide a suggestion of how to understand the terms "extraordinary" and "ordinary" within Sagan's principle on the basis of prior likelihoods and the cost of mistakes: An extraordinary claim is one likely to cause high costs in the case of a mistake. In the second section, I review a recent debate concerning the sources of evidence for the study of well-studied languages (GIBSON & FEDORENKO, 2010; SPROUSE & ALMEIDA, 2013), and show that the outcome of the debate has essentially been Sagan's principle. In the second part, I review recent contributions on the methodology of the study of less well studied languages, especially by DIXON (2007) and MATTHEWSON (2004), and point out some weaknesses of Dixon's text-only method. But while I agree with Matthewson's inclusive approach for most cases, I show that there are cases where one should call upon experimental evidence: specifically, I discuss MATTHEWSON'S own (2006) work on

presuppositionality in Salish as a case in point. Instead of Dixon and Matthewson fieldwork specific methodologies, I propose that Sagan's principle should be applied as a guideline for the study of less well studied languages as well. This entails that both evidence from traditional fieldwork techniques as well as evidence from experimental techniques have a role to play. For many questions, though, a combination of established fieldwork techniques with proper documentation using modern recording technologies is most appropriate as the main source of evidence. In the final section, I conclude with some practical suggestions for when to incorporate formal experiments to gather quantitative evidence in syntactic and semantic fieldwork.

1 A Philosophy of Science Briefing

1.1 Sagan's Principle and its Implicature

The first part of my title is a quote from the physicist SAGAN (1980): "Extraordinary claims require extraordinary evidence." In the title, I furthermore articulate an implicature of Sagan's principle, namely that ordinary claims require only ordinary evidence. In this section, I argue that the Sagan's principle as well the implicature I added are an established tenet of the philosophy of science dating back to at least HUME (1748). I furthermore consider how it can be applied to evidence in linguistics.

I first show that Sagan's principle is actually rooted deeply in the history of science. I quote Sagan's principle here from a science TV program Sagan appeared in. While Sagan's phrasing is widely quoted and original to Sagan, it is also well-known that the underlying principle is much older than Sagan's formulation of it. Two much older formulations

of essentially the same principle are due to HUME (1748) and LAPLACE (1814). Two relevant quotes from section 10 of HUME'S (1748) book are "A wise man ... proportions his belief to the evidence" and "No testimony is sufficient to establish a miracle, unless the testimony be of such a kind that its falsehood would be more miraculous than the fact which it endeavors to establish." Hume calls the latter quote a "general maxim worthy of our attention", so we might also use the term Hume's maxim instead of Sagan's principle. A second, early relevant quote is the following from the French scientist LAPLACE (1814: p. 50): "Qu'it ne serait pas philosphique de nier les phénomènes, uniquement parce qu'ils sont inexplicable dans l'état actuel de nos connaissances. Seulement, nous devons les examiner avec une attention d'autant plus scrupuleuse, qu'il paraît plus difficile de les admettre." ("It would not be philosophy to deny phenoma solely because they are inexplicable according to the present state of knowledge. But we ought to examine them with an attention all the more scrupulous as it appears more difficult to admit them.") WIKIPEDIA (2014b) reports that FLOURNOY (1899) reformulated Laplace's principle as: "The weight of the evidence should be proportioned to the strangeness of the facts.", which might have inspired the more modern formulation of Sagan. The quotations show that Sagan's principle is an old principle of science.

Consider now the implicature of Sagan's principle that ordinary claims require only ordinary evidence. In Hume's and Sagan's formulations, the implicature I added in the title is not made explicit. Conditionals "If p then q" are well known in the linguistic literature to trigger an implicature "If not p then not q", also referred to as conditional perfection (GEIS & ZWICKY, 1971 and others). For example the recommendation "If it's raining, you should take an umbrella" implicates that if it's not raining, you shouldn't take an umbrella. Sagan's principle is just an elegant formulation of the conditional "if you make an extraordinary claim,

you must present extraordinary evidence for it." So, it implicates that ordinary claims require only ordinary evidence. The same holds for the second quote from Hume above: The statement that no evidence except for a very strong kind is sufficient to establish a miracle, implicates that there is evidence that's less than very strong, but sufficient to establish a non-miracle. Actually both HUME (1748) and SAGAN (1980) had a reason to leave the implicature implicit: The quotes of Hume's are from a chapter on miracles, so Hume was focussing on the case of extraordinary claims. A similarly reason applies to Sagan's quote. The Sagan quote is inspired by the rather similar phrase "An extraordinary claim requires extraordinary proof" by the sociologist TRUZZI (1978) according to WIKIPEDIA (2014a). While I have no opinion on whether Sagan was actually quoting Truzzi, it is of interest to the present argument that both Sagan and Truzzi were fellow founders of the Committee for the Scientific Investigation of Claims of the Paranormal. Given this fact, it makes sense for both of them to omit the implicature concerning ordinary claims because all claims of paranormality are by definition extraordinary. The first Hume quote and also Laplace's and Flournoy's formulations are explicit about the implicature since they speak of a proportional relation between the evidence furnished and the claim that one is sought to establish.

In sum, I showed in this section that Sagan's principle including its implicature are established principles in the philosophy of sciences. However, we still need to determine how to apply the principle to linguistic claims and evidence. To be able to do so, we need to understand the adjectives "ordinary" and "extraordinary" as applied to linguistic claims and evidence. In this next section, I suggest a general Bayesian understanding of both terms and consider specifically what this entails for linguistic evidence.

1.2 What's Extraordinary? The Cost of Errors

To derive any consequences from Sagan's principle, we need to understand what constitutes and ordinary vs. an extraordinary claim and similarly what constitutes ordinary and extraordinary evidence. Hume's term "miracle" indicates that extraordinary claims are those that we believe to be highly unlikely. But, I think we also intuitively understand that the cost of a potential error affects the quality of the evidence we desire: Before you go on an overseas trip, you might check several times that you have your passport with you. But you are more likely to not check at all that you packed your tooth brush. That this behavior is rational at least if you re are anything as forgetful as me, follows from the Bayesian computation linked to the cost of error. Assume that your equally likely to not have packed your passport or your toothbrush. But the cost of not having your passport is substantial: you might not be able to board your flight without it. And since even the memory that you checked your passport 20 minutes ago might be mistaken, it makes sense to expend the energy to check again just to be sure you avoid the substantial cost of error. By comparison, the cost of not having your toothbrush is a lot smaller, so checking whether you have it on you is uneconomical -- the error is too inconsequential to be worth the cost. The example shows to factors that play a role: the cost of the test and the cost of an error. A further factor that plays a role is the reliability of a test: Staying with the example, you might check for your passport either by quickly feeling through the outside of your bag that it contains a passport sized printed document or you might do a more elaborate, but more reliable check: open your bag, take out the passport, open it, and check the name and validity of the passport.

The statisticians NEYMAN and PEARSON (1928) introduced the discussion of two different type of error types in testing a hypothesis. A type I error occurs when the test comes out in favor of a hypothesis

that's actually false, while a type II error occurs when the test results speaks against a hypothesis that's actually true. The two types of errors might cause different amounts of damage. The calculations involved in this type of scenario well understood in the context of medical tests and is discussed in conjunctions with the concepts of sensitivity and specificity of tests. For our purposes a rough understanding is sufficient. Consider the case of a medical test for a condition that has no other symptoms. The test itself has a specific cost to it both in terms of the cost of carrying out the test and in terms of the suffering the test itself causes. Assume we furthermore know that a specific percentage of a specific population has the condition in question. Finally we know the rates of the two errors: For the people that actually have the medical condition in question, the test gives a specific rate of Type II errors or false negatives. And for the people that actually don't have the medical condition, the test gives a specific rate of Type I errors or false positives. On the other hand, the true positives benefit from treatment, while for the true negatives only the cost of the test itself is incurred. The framework of utility analysis of von NEUMANN & MORGENSTERN (1944) implies that the test is only useful if the cost of the test to an individual is smaller than the potential benefit that individual draws from each of the four possible outcomes times the possibility of belonging to each of the four groups. And if we have to decide between two possible tests, we would want the cost increase the more expensive test causes to be smaller than the total additional benefit an individual gets from the pricier test as compared to the cheaper test. Schematically, a test with many false positives is more acceptable when the likelihood of an actual positive increases or the cost done by treating a false positive decreases. And a test with many false negatives is more acceptable when the actual rate of positives is decreased or when the cost a false negative is decreased. In medical testing all the precise costs may be known at least in approximation: the cost of the test and the treatment, the cost

reduction of successful treatment compared to not giving treatment to affected individuals and also the cost of the possible side effects of treating one of the false positives. In linguistic examples, the cost of a test is generally known in approximation: it's roughly the cost of the required field work. But the benefit of a correct linguistic theory or the cost of an incorrect linguistic theory are not known. So we can't really determine what rate of Type I and Type II errors we should tolerate.

In behavioral psychology, researchers have agreed to tolerate 5% of Type I and 20% of Type II errors. These values apply specifically to a scenario of testing a tested hypothesis that two groups of measurements are drawn from different populations vs. the null hypothesis that the two are drawn from two randomly chosen groups of the same population. The 5% and 20% rates correspond to the assumption that there is no prior bias as to whether the test hypothesis is correct or not, and but that the cost of publishing a false positive is greater than the cost of not publishing a false negative (in terms of cost to the field of psychology, not to the individual researcher). There are many scenarios in linguistics and also in linguistic fieldwork where this kind of set-up is applicable. For example, we may compare whether males and females use an overt first person pronoun, starting from the null hypothesis that there is no sex difference. In such a case the run-of-the-mill psychological method can be applied to possibly show that there is a sex difference. However, in linguistic work and especially in linguistic field-work, situations are common where the psychological method isn't applicable. Consider for example the situation of a field-worker who wants to determine the word used to describe *rabbit* in the language of an indigenous group. QUINE (1957) argues that it is close to impossible to conduct even such a simple task while strictly applying the method of behavioral psychology. The field-worker would need to set-up a series of controlled experiments where indigenous speakers observe rabbits vs. some other

stimuli to determine that the presence of rabbits triggers a different word from that of for example foxes and thereby shoot down the `null-hypothesis' that the words for rabbits and foxes is the same in the indigenous language. Quine concludes from this argument that field-work can never yield determinate results and more generally that translation is indeterminate, but this result derives in large part from Quine's strict adherence to the behavioral method. Some residual uncertainty is common to all scientific endeavors, of course. But this is better discussed for medical research, nuclear physics, or the theory of evolution rather than linguistic fieldwork.. At the same time, the field-worker has a justified prior belief that the perceptual organs and the mental faculties of the indigenous groups do not substantially from those of other humans. Why then should the field-worker start with the null-hypothesis that the indigenous group should lack a term for *rabbit* if rabbits occur frequently in the environment the indigenous group occupies? A different kind of starting point would be the hypothesis that the indigenous language is not substantially different from other previously studied languages including the well-studied European and East-Asian languages except for the phonetic content of the lexical items. Under this perspective the Type I and II errors are the opposite from that of the Quinean perspective. So if we accept different rates of the two types of errors, we arrive at different outcomes. Assume we accept 20% of Type II errors. Then, on the Quinean approach up to 20% of the claims of the form "this indigenous group doesn't have a word for X" would be false, while on the latter approach up to 20% of the claims of the form "this indigenous group has a word for X just like European languages" would be false.

The previous paragraph I argued that the choice of a null hypothesis is by no means obvious for linguistic research. One approach might lead to over-exoticizing the language under study, the other to over-

Europeanizing it. Over-exoticizing arises as follows: If field-worker was to generally start from the null hypothesis that an indigenous languages lacks a specific distinction like that between rabbits and foxes and we accept a high rate of false negatives, the indigenous language would end up being described as lacking many distinctions better studied languages draw. But, over-Europeanizing comes about when the field-worker starts with the assumption that the indigenous language is similar to well-studied languages. Then any false negative corresponds to a claim that the language in question has a property of some well-studied language. Regardless of approach, errors are of course unwanted and the expectation is that attempts at replication of a result are going to eliminate errors over time. But some rate of error is unavoidable in any scientific endeavor, and given that, the error of over-Europeanizing is the less dangerous type of error. In the following, I'll call this the comparison-based approach: the term *Over-Europeanizing* brings with it connotations of neo-colonialism and cultural imperialism, as well as memories of Latinizing the description of some European languages by religious scholars and translators of the Bible. But in current linguistic work, a substantial variety of languages is quite well-described so the connotations mentioned don't apply. The body of current grammatical description is certainly still dominated by languages from the Indo-European family, but detailed descriptions of several East-Asian languages, languages from other families spoken in Europe and its periphery (Finno-Ugric, Turkic, Semitic, Basque), and substantial amount of grammatical description of other languages from all over the planet. Therefore the null hypothesis would in most cases need to be specified as which is the language of comparison: this phenomenon in language X is like a specific phenomenon in the specific better studied language Y, where the field-worker would need to take care to determine the right comparison language Y. In addition, the field-worker would also need to identify the phonetic content of the lexical items. So,

there are sufficient checks built into the comparison-based approach to ensure that erroneous claims of the type that language X is like some well-studied language in some respect are not damagingly frequent. Furthermore claims of the type language X is like language Y rarely excite great interest, and therefore the greater burden of proof should be placed on the researcher claiming that language X and language Y differ. The over-exoticizing approach, in my view, is more severely handicapped: it seems to start from the presumption that the humanity of the indigenous group is in doubt and ends up all too frequently making false pronouncements of language X lacking some property simply on the basis of the field-worker not having found positive evidence for the property in question. Given that claims of this type incur great interest, the field shouldn't rely on an approach that generates a high number of false negatives of this type.

In sum, I have argued that claims in fieldwork on indigenous language should generally be taken to be ordinary if they closely correspond to generalization established on the basis of better studied languages. And one property of an extra-ordinary claim is that it claims that an indigenous language diverges from the grammatical properties of better studied languages. The resulting picture is different from one where it's assumed as null-hypothesis that some grammatical factor always plays no role in the indigenous language. In addition the degree of extra-ordinariness depends on the hypothetical cost caused by an erroneous claim (a false positive) and that of an erroneous rejection of the same claim (a false negative). But this cost can not even be estimated at this point and the actual decisions depend largely on the social consensus of the researchers in the field -- given what kind of evidence are others willing to revise their theoretical models to incorporate the new claim. In the following sections, I attempt to derive more practical consequences out of these general philosophical principles.

2 Current Linguistic Methodology

2.1 Evidence from Well-Studied Languages

This section summarizes the current state of a debate about which is the most suitable method to gather acceptability judgments in one of the best studied languages there is: English (GIBSON and FEDORENKO 2010, 2013, SPROUSE and ALMEIDA 2013, SPROUSE et al. 2013). The recent debate began with a broad accusation of sloppiness against research using traditional "armchair" methods in syntactic and semantic research. But at least at this point, the result of the debate is that quantitative evidence from formal experiments offers no better validity than evidence from the traditional "armchair" method. This is an important result to keep in mind also for linguistic work on less well-studied languages where also usually quantitative evidence is not collected.

GIBSON and FEDORENKO (2010, 2013) accuse the fields of syntax and semantics of being open to researchers own cognitive bias and specifically a confirmation bias in favor the researchers own proposal. To support this claim they cite a couple of selected, individual cases of judgment data from the published literature that Gibson and Fedorenko failed to reproduce in quantitative studies involving multiple test conditions and multiple speakers. Gibson and Fedorenko therefore call for the widespread adoption of quantitative research methods for syntax and semantics. Given the advances in software technology to conduct judgment elicitation over internet platforms such as Amazon Mechanical Turk, they claim that the cost both in terms of researchers time and payment of participant would be worth the putative gain in accuracy. SPROUSE and ALMEIDA (2013) and SPROUSE et al. (2013), however, show that Gibson and Fedorenko themselves are guilty of a

violation of one basic of method of quantitative research: they don't consider a random sample of data in their evaluation of the "armchair" research method, but instead focus on a few selected cases of data that were already known to be controversial. SPROUSE et al. (2013) present an evaluation of the "armchair" method based on a randomized selection of data from journal articles. They report that 95% of the contrasts in acceptability they tested are confirmed by quantitative measurement using Amazon Mechanical Turk. This means that the "armchair" method is at least comparable concerning the number of false positives it results in than the standard method of behavioral psychology. The failed confirmation of 5% of the contrasts in the quantitative trials may signal that there is indeed some small mismatch, but at this point it remains open as to whether these are false positives of the "armchair" method or false negatives of the quantitative method. Overall this result entails that existing results and methods in the field are not undermined by the wider availability of experimental methods, but experimental methods are still important as they allow research on many questions that couldn't be addressed by the "armchair" method. Also for surprising, extraordinary new claims there could be an advantage to providing stronger evidence as expected by Sagan's principle.

The finding of SPROUSE et al. (2013) is important for the fieldwork since it shows that currently most contrasts of acceptability relevant to linguistic theory are such strong effects that they can be reliably judged without resort to quantitative methods at least by trained linguists. There are some difference though that remain to be investigated: In most fieldwork situations, the bulk of data will be collected not from trained linguists, but from between one and a couple of language consultants. Also while most interested researcher (e.g. the reviewers of a paper) can readily attempt to reproduce English judgments, this is in most cases impossible for indigenous languages. In the following section, I consider existing recommendations for the methodology of fieldwork

and the implications of the wider availability of documentation and experimental techniques.

2.2 Evidence for Less Well-Studied Languages

In this section, I first focus on two recent contributions on the methodology of field work on less widely spoken languages by two prominent researchers in the field: MATTHEWSON (2004) and DIXON (2007). The two approaches represent two opposite ends of a spectrum of opinion (and therefore are exemplary for other's views of others in the field): while Dixon urges an almost exclusive focus on the collection of texts in the target language, Matthewson advocates in addition the use of elicitation, of translation and use of a contact language in the field. But neither of the two explicitly addresses the issue of formal experiments. The goal of the section is to defend briefly again the liberal view of Matthewson, but in addition to indicate some space where formal experiments should be added to the fieldworkers inventory.

First, consider the recommendations DIXON (2007) offers. While Dixon's paper contains more general advice on the practical aspects of fieldwork, a substantial part of his paper concerns methodology. Here, the focus on collecting texts is very explicit in Dixon's section 9 on "what to do". Dixon focuses on three tasks: beginning to speak the language, compiling a dictionary, and recording and analyzing texts. The list doesn't mention grammatical elicitation, and Dixon states at the end of the section (p. 23) that grammatical elicitation "should play *no role whatsoever* in linguistic fieldwork" (emphasis in original). Also on p. 22, Dixon writes that "the only way to understand the grammatical structure of a language is to analyse recorded texts in that language." Furthermore, in a later section on "what not to do" (p. 27), Dixon reiterates that controlled elicitation shouldn't be pursued.

Dixon's view as far as I can gather is extremist and Matthewson mentions many researchers who have taken a different stand. Nevertheless the extremist texts-only view still remains influential in the field. But there are many problems with the exclusive focus on collecting texts. One general problem of corpus based linguistics is that it neither obtains ungrammatical sentences nor sentences that are false in a specific scenario. Given the important role these types of data play in linguistic analysis, the text-only view is prone to a large Type II error: not finding data that would actually show interesting distinctions. Consider briefly two examples that illustrate the error-proneness of a text-only strategy. The first example is actually from spoken English. Spoken English isn't a bad proxy for an indigenous language in a field-work scenario which likely doesn't have a written form, but at the same time English is well studied. THOMPSON (2002) looks at evidence for complement clauses in spoken English corpora, and concludes that rather than complement clauses, spoken English only allows unembedded declaratives accompanied by an evidential phrase. To support her claim, Thompson extracted a sample of 452 complement-taking predicates from a corpus of spoken English and analyzed the structure and discourse contribution of each item in detail. NEWMEYER (2010) points out that Thompson committed a Type II error: concluding from the lack of evidence, that something doesn't exist. Specifically, Newmeyer looks at a much larger corpus of English than Thompson did: 170 Megabytes of text data, and finds numerous different types of evidence for the existence of complement clauses in English. What is instructive here is the amount of text required to avoid a Type II error: For which indigenous language has anybody gathered and transcribed 170 Megabytes of data? If one page of text corresponds to 500 Bytes (characters), then 170 Megabytes correspond to 340 thousand pages of text: Dixon's text-only approach requires the field-worker to gather roughly two bookshelves full of transcribed stories required to determine whether a language has

complement clauses or not. At the same time, the text-only approach is not immune to errors of the Type I type. In fact, once a linguist is actually immersed in a collection of texts and their translations, it seems to become difficult to discern specific surprising properties of the language under investigation. An example indicating that type of failure to see an obvious linguistic difference to widely spoken languages in a text that was widely studied both in original and in translation concerns Homeric Greek. As DEUTSCHER (2011) renarrates, the fact that the color categories of Homeric Greek don't correspond to color categories of English, German, French and other modern European languages was only pointed out by GLADSTONE (1858: 457-499). Gladstone slightly overinterpreted the data (He assumed color vision was different at Homer's time), but given the modern knowledge of cross-linguistic variation in color terms Gladstone's basic observation was essentially correct, though overlooked by almost all others studying Homer's writing. I don't think we can be confident that similar oversights can be ruled out if linguistics was to rely entirely on the text-only method. Two factors may lead field workers to overlook interesting grammatical properties in texts: For one, since the gathered texts are usually stories, field workers may attribute unusual properties to metaphoric use or poetic language as was the case for Homeric Greek, but one Gladstone argues to be incorrect. Secondly, relevant data in texts may be spread out very thinly over different stories, and may only be convincing when arranged into one paradigm. This probably was less of a factor in the case of Homer since many researchers intensively studied his writing, but it is a significant concern for texts gathered in fieldwork. Needless to say the very successful typological research on color terms by BERLIN & KAY (1969) didn't rely on text corpora at all.

In sum, the text-only approach is prone to a large number of Type II errors and even then may not generally lead to a full view of the

language under investigation. A third concern about the approach is that it might lead field workers to not document a large part of what they do. I have always found it very natural to construct an example sentence in a language I'm interested in and then ask a native speaker of whether it's grammatical and what it means. I can't imagine that other linguists differ in this respect, even those who subscribe to the text-only view. In fact, Dixon's manual urges the field-worker to learn the indigenous language and try to use it in conversation. Furthermore, he says it's important "to encourage people to correct all your mistakes" (p. 20), which is very close to judgment elicitation: tell me whether what I say is grammatically correct and whether it's true in this scenario. The major difference here is that Dixon doesn't exhort the field-researcher to document exchanges of this type, while I consider this at least as important to document such judgement-elicitation sessions as stories and other texts: Only if other researchers have access to all the data a field-worker's conclusions are based on (including the ungrammatical sentences), can they evaluate the arguments and it is this independent scrutiny that underpins progress in the field. Furthermore, in such judgment elicitation sessions all important grammatical sentences should be recorded from at least one native speaker.

MATTHEWSON (2004) is primarily concerned with semantic fieldwork. The distinctions between syntactic, semantic, and pragmatic field-work aren't generally easy to draw though, and Matthewson's paper contains a lot of insights relevant to these other subfields. Matthewson like me takes a strong stance against the text-only approach. She in particular argues that the use of a contact language (she uses the term "meta-language") and of translations in elicitation need to be handled with care, but need not be detrimental. One recent example of the latter from work on Matses I was involved in is presented by MUNRO et al. (2012): We investigated the claim that Matses doesn't have a form of

speech report corresponding exactly what is indirect speech in English and also in Spanish. As part of a controlled elicitation experiment, we asked Matses speakers who also spoke Spanish to translate sentences with indirect speech from Spanish into Matses. In this experiment, we might have easily ended up with data that show transfer effects from Spanish into Matses. However, actually eight of the nine speakers gave Matses responses that fully corresponded to the claim that Matses only has forms similar to direct speech in Spanish. The ninth speaker, who did show a transfer from Spanish to Matses, was working as a Spanish teacher. So, the example illustrates the potential for transfer effects involved in translation tasks, but at the same time the evidence obtained is actually revealing when speakers overcome the potential for grammatical transfer inherent in translation tasks. Overall, Matthewson's liberal approach is on the right track in attempting to strike a balance between attempting to block false positives, but at the same time not impose unreasonable methodological barriers that impede progress and cause a large number of false negatives simply because research methods commonly used for widely spoken languages are banned for indigenous languages.

One topic Matthewson doesn't address is the use of formal experiments in fieldwork. As I discussed in the previous section, there is no general, precise answer to this question until we know the cost of having different types errors in our linguistic theory. If one is too skeptical, one would end up spending a lot of field work time and effort for results that could've been obtained with equal accuracy in an easier way. The satisfactory reliability of the "armchair" method shows that not every formal experiment is warranted. On the other hand, if one doesn't undertake the effort of a formal experiment when it's due, one may easily miss an opportunity to convince the community of an observation that is extraordinary in Sagan's sense. There can't be a general answer to when formal experiments could add to field-work

results, even assuming that the experiment if properly designed and carried out. One example where a non-experimental fieldwork result has been ignored and this lack of reception is possibly due to non-experimental nature is provided by Matthewson's own work on Lillooet Salish. MATTHEWSON (2006) proposes on the basis of fieldwork that St'át'imcets lacks presuppositions of the type that English has which place a requirement on the common ground (STALNAKER, 1973 among others). She proposes that instead St'át'imcets can mark content as objective propositional content in the sense of GAUKER (1998). To argue for this parameter Matthewson cites, on the one hand, unpublished experimental work by CONTI (1999), which I couldn't access, and published work by herself (MATTHEWSON et al. 2001) on English, and, on the other hand, reports on an informal experiment conducted in the field for St'át'imcets. MATHEWSON (2001, 2006) relies on the frequency of presupposition challenging responses to test for presuppositionality. Specifically, she tested adult English speakers on the presupposition of (1) that there is in elephant in your hair isn't satisfied, and found that 62% of them challenge (1) with a phrase like "What elephant?" or similar.

(1) Did you get **the** elephant out of your hair?

In 2006, Matthewson claims that St'át'imcets adults don't challenge what might appear to be presuppositions in the same way. For example, she reports on presenting the sentence in (2) to one of her consultants. The English translation of (2) presupposes that some other person's being in jail was mentioned before, but Matthewson reports that her St'át'imcets informant didn't challenge this response, but only asked what Lisa did to land her in jail. Matthweson takes this to be evidence that "t'it" in (2) doesn't trigger a presupposition of the same type as English presuppositions, though it has the same semantic content.

(2) wá7 **t'it** l-ti gélgel-a tsitcw k Lisa
be **also** in-DET strong-DET house DET Lisa
'Lisa is also in jail.'

As far as I know Matthewson's claim has been largely ignored in the field. One current theme in work on presupposition (e.g. ABRUSAN 2011) is to attempt to derive that some aspects of content must be presuppositional from general pragmatic and semantic principles. But if MATTHEWSON (2006) is correct, that enterprise would be futile since the parametric difference between English and St'át'imcets shows that the presuppositionality of some content is an arbitrary feature of a specific language like English and other languages can differ. I think the reason Matthewson's work has been ignored is that the evidence she present has not been as extraordinary as the claim Matthewson is making, and she should've done a more formal experiment on the matter. That Matthewson's claim is extraordinarily surprising is in this case clear: MATTHEWSON (2006) writes herself that her claim is "somewhat radical" (p. 63). HER 2006 generalization also differs from HER previous 1998 work, where she writes (p. 116) that "the lexical item corresponding to English 'too' induces presuppositions"[2].

I conclude therefore that Matthewson should have done formal experimental work to corroborate the central claims of her 2006 work. Specifically, one kind of study that comes to mind is the following: Matthewson compares the performance of English adults on (1) with that of St'át'imcets adults on (2). But neither the two groups nor the two sentences are very similar: As for the two groups, MATTHEWSON (2006) writes that her consultants "My relationship with the consultants from whom data were obtained is a friendly one, and I have known each of

[2] MATTHEWSON (1998) proposes that determiners in Salish languages and specifically St'át'imcets cannot be presuppositional, consistent with the 2006 claim. Furthermore, she does preface the discussion I quote from in the text with the proviso that the full investigation of the relevant claims is beyond the scope of her 1998 work.

the consultants for between 12 and 14 years." (p. 68). MATTHEWSON et al. (2001) don't report who the English adult subjects were that gave the 62% challenge-responses, but quite likely they were undergraduate students at the University of Massachusetts. If this is correct, there appear to be substantial differences in age, level of education, and familiarity with the experimenter between the two groups. As for the sentences, a more appropriate comparison would've been between (2) and its English translation. If a simple formal experiment controlling these two factors better corroborated them, it would certainly have brought broader recognition to MATTHEWSON'S (2006) results.

Overall though the observational method has served the field quite well. Though some regard Galileo's experiments as the starting point of modern science, many big discoveries in science weren't based on experiments, but only on close observation of nature. Consider just one of the most influential modern scientific theories: Darwin's theory of evolution. Almost the entire body of evidence for evolution that Darwin based his theory comes from observations. Darwin wasn't opposed to experiments -- late in his life, he proved experimentally that earthworms improve the fertility of soil --, but Darwin did not waste any time on establishing obvious facts underpinning his account of evolution such as the anatomy of the Galapagos fauna, and at his time of writing also didn't have the methods at his disposal to seek experimental confirmation of evolution. Linguists too need to develop a taste for when quantitative data are useful, and when they are an impediment to progress.

3 Conclusion: When to formally experiment in the field?

To conclude, I take stock and attempt to derive some practical recommendations. It is impossible, though, to derive a precise general recommendation from the above considerations as to when to do a

formal experiment and when it would be a waste of time. Clear cases of former are any cases where experiments are also called for for well-studied languages: cases where the data are subtle, cases where additional measurements such as timing data or neurological data are expected to be revelatory, and cases where groups other than adult informants are under investigation. Clear cases of the latter are data that trust-worthy language consultants judge to be clear and that conform to patterns of a better studied language. That leaves a large area where it is up to the individual researchers judgment whether experiments are expected to add to the reliability of the findings. But I think there is enough motivation to suggest that researchers engaged in fieldwork should at least consider and acquire the ability to perform formal, quantitative experiments. Especially this should be the case in situations where formal experiments can be incorporated into the fieldwork situation without being taking away a large amount of time from other methods like judgment elicitation and story elicitation.

Compared to the investigation of well-studied languages the situation for field-work is different in several ways. One recent technological advance in the case of well-studied languages has been the availability of the internet based platforms that allow researchers to conduct trials, especially Amazon's Mechanical Turk (a point of GIBSON & FEDORENKO'S 2010 paper that I discussed above in section 2.1). The internet based methods make it possible to conduct quantitative trials for researchers that don't have access to lab space and research assistants that gather judgment data from a large group of English speakers. But, speakers of any indigenous language are unlikely to be available on Mechanical Turk -- even for German and Japanese I found it impossible to get more than about 40 native speakers in a study conducted in 2011. But integrating formal experiments into fieldwork also benefits in some ways from recent technological progress and furthermore offers some

distinct advantages. The two ways technological progress make it easier to conduct formal experiments in the field are the following: For one, smaller and cheaper equipment make it possible to create and manipulate recordings and other data in the field to set up an experiment. Secondly, freely available, public domain software tools facilitate both the creation of stimuli and analysis of data from formal experiments. For the analysis, especially the Praat software for phonetic analysis and the R software for statistical analysis and the creation of graphs can nowadays replace expensive commercial software such as SPSS for research purposes.

Still it costs time to prepare and conduct formal experiments to gather quantitative data in the field. But there a number of important reasons to consider doing so in selected circumstances. The typical scenario I have in mind here is one where initial fieldwork using judgment elicitation has already provided evidence in favor of the conclusions that a formal experiment then attempts to corroborate. Of course there are also cases where judgment elicitation isn't useful in the first place as with questions of traditional psycholinguistics and experiments are needed to establish any useable evidence. But even in cases where language consultants judgment provide some evidence, there may be reasons to conduct in addition a formal experiment. The foremost reason is that the evidence and the conclusions drawn from it are surprising, for example they may contradict apparently well-established generalizations in linguistics. If this is the case, the simplest possible experiment would be to independently question several additional members of the community. After all, if all five out of five individual agree that a contrast goes the same direction, this establishes statistical significance by the binomial test. A second good reason may be that in situations of language endangerment the evidence might not be available later. So it may be the last chance to document any property of such a language with greater reliability. A related third reason is that in situations where an indigenous language

is not the main daily language of most members of a community anymore, individual consultants judgement may be more uncertain than otherwise, or it may be desirable to determine whether all members of the community share the relevant judgment. For example, in work of my own with the Teiwa in Indonesia (Kratochvil, Hollebrandse, and Sauerland, in progress), we primarily worked with younger speakers as consultants. However, younger speakers were all literate in Indonesian and we turned to experimental techniques to determine whether older, illiterate speakers shared the relevant judgments. A fourth advantage of integrating some experimental work in field situations is that it can involve all members of the community, while typically one works every day with the some preferred consultants in judgment elicitation who gain some proficiency, and also not all community members volunteer to tell stories to be recorded. In this situation, enrolling all comers as participants in an experiment and providing some appropriate compensation for the effort is a way to engage the whole community in the study and receive some immediate benefit. For this, it is helpful if the experiment is not too hard, but rather fun for the participants, so I offer some simple methods in the following.

For work in syntax and semantics, simple methods are focused on the task of either judging that one sentence sounds better (i.e. more grammatical) than another, or that one sentence is more acceptable in a specific situation. It is also often helpful to look at work done in language acquisition research with children since materials for children must be designed to be engaging. Of course, one shouldn't overdo this: infantilizing is no better received by members of an indigenous community than by adults elsewhere. It also is helpful to ask the community or the consultants one works with closely for suggestions on how to do this. My own experience comes from work on Teiwa mentioned above, Matses (MUNROE et al. 2012), and Pirahã (Sauerland, to appear). The methods

that are most easily and broadly applicable: we essentially compiled a list of testable examples working with a few primary informants. The list consisted primarily of questions of the type, can this sentence be used in this scenario? After the list was done, we went around with it and asked other speakers the questions on the list, and just reported their responses. In addition we did another experiment involving translations from Spanish into Matses. In the case of Teiwa and Pirahã, the goal was different from the one in Matses. For both languages, it was unclear if specific scenarios (namely, false belief scenarios) could be described in the language at all. In this case, we followed language acquisition research by using targeted elicitation: We created relevant scenarios and asked speakers to report relevant aspects of that scenario. So this was spontaneous production in a controlled situation. If it works, the results from this method provide strong evidence for the existence of the specific structures speakers use in this scenario since speakers produce them spontaneously. But the work required especially for the evaluation is much greater, than in acceptability or comprehension studies, and additional comprehension experiments seemed necessary to me in both of my cases to confirm the interpretation. This examples illustrate that, if formal experiments are warranted in the first place, the goals determine the best method and setup.

References

ABRUSAN, Marta. **Predicting the presuppositions of soft triggers**. Linguistics and Philosophy 34: 491–535. 2011.

BERLIN, Brent & Paul Kay. **Basic color terms:** Their universality and evolution. University of California Press. 1969.

CONTI, Rachel. **Presuppositions of the.** Ms., University of Massachusetts, Amherst. 1999.

DEUTSCHER, Guy. **Through the language glass:** Why the world looks different in other languages. Arrow Books. 2011.

DIXON, R. M. W. **Field linguistics: a minor manual**. Sprachtypologie und Universalien Forschung (STUF). 60: 12–31. 2007.

GLADSTONE, William E. **Studies on Homer and the Homeric Age I.** Oxford, United Kingdom: Oxford University Press. 1858.

HUME, David. **An Enquiry concerning Human Understanding**. 1748.

FLOURNOY, Théodore. **Des Indes à la Planète Mars:** Étude sur un cas de Somnambulisme avec Glossolalie. 1899.

GAUKER, Christopher. **What is a context of utterance?** Philosophical Studies 91. 149–172. 1998.

GIBSON, Edward & Evelina Fedorenko. **Weak quantitative standards in linguistics research**. Trends in Cognitive Science 14. 233–234. 2010.

Gibson, Edward & Evelina Fedorenko. **The need for quantitative methods in syntax and semantics research**. Language and Cognitive Processes 28, 88:124. DOI: 10.1080/01690965.2010.515080. 2013.

KRATOCHVIL, Hollebrandse, and Sauerland. in progress. **Complexity before Evolution:** Complement Clauses in Teiwa.

LAPLACE, Pierre Simon. **Essai philosophique sur les probabilités**. 1814.

MATTHEWSON, Lisa. **Determiner Systems and Quantificational Strategies:** Evidence From Salish, Holland Academic Graphics, The Hague. 1998.

MATTHEWSON, Lisa, Timothy Bryant and Tom Roeper. **A Salish stage in the acquisition of English determiners:** Unfamiliar 'definites'. In The Proceedings of SULA. GLSA, University of Massachusetts, Amherst. 2001.

MATTHEWSON, Lisa. **On the methodology of semantic fieldwork**. International Journal of American Linguistics 70. 369–415. 2004.

MUNRO, Robert, Rainer Ludwig, Uli Sauerland & David W. Fleck. **Reported speech in Matses:** Perspective persistence and evidential narratives. International Journal of American Linguistics 78. 41–75. 2012.

NEWMEYER, Frederick J. **What conversational English tells us about the nature of grammar:** A critique of Thompson's analysis of object complements. In Kasper Boye & Elisabeth Engberg-Pedersen (eds.), Usage and structure: A festschrift for Peter Harder, 3–43. Berlin: Mouton de Gruyter. 2010.

NEYMAN, Jerzy & Egon S. Pearson. **"On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference, Part I".** Joint Statistical Papers. Cambridge University Press. pp. 1–66. 1928 [1966].

QUINE, Willard van Orman. 1960. **Word and Object**. Cambridge, Mass.: MIT Press.

SAUERLAND, Uli. to appear. **False Speech Reports in Pirahã:** A Comprehension Experiment.

SAGAN, Carl. **"Cosmos:** A Personal Voyage", Episode 12, "Encylopedia Galactica", 1:10 min. 1980.

SPROUSE, Jon & Diogo Almeida. **The empirical status of data in syntax:** A reply to Gibson and Fedorenko. Language and Cognitive Processes 28. 222–228. 2013.

STALNAKER, Robert. **Presuppositions.** Journal of Philosophical Logic 2. 447–457. 1973.

THOMPSON, Sandra A. **'Object complements' and conversation:** Towards a realistic account. Studies in Language 26. 125–164. 2002.

TRUZZI, Marcello. **"On the Extraordinary:** An Attempt at Clarification." Zetetic Scholar, Vol. 1, p. 11. 1978

WIKIPEDIA. 2014a. **"Marcello Truzzi."** http://en.wikipedia.org/wiki/Marcello_Truzzi, accessed Aug. 2, 2014

WIKIPEDIA. 2014b. **"On Miracles."** http://en.wikipedia.org/wiki/Of_Miracles, accessed aug. 2, 2014.