Journal of Universal Language 4
March 2003, 33-74

Languages and Universals

David Odden
Ohio State University

Abstract

A central question of linguistic research for nearly a half century has been whether there are properties universal to all human languages. There are many ways of conceptualizing linguistic universals, but at the core, the idea of linguistic universals asserts that some observed properties of human language are arbitrary—the fact that "dog" is pronounced [d g] in English, [kæ] in Korean and [mbwa] in Swahili—but a number of facts of language are not random and arbitrary. If there are non-random properties of linguistic structure, questions arise regarding those properties: what are they; how do we identify them; why do they exist?

Keywords: observational universals, word-order universals, distribution of language

1. Types of Universals

One example of a universal property is the Coordinate Structure Constraint proposed in Ross (1967: 89), which states "In a coordi-

nate structure, no conjunct may be moved, nor may any element contained in a conjunct be moved out of that conjunct". This principle explains why English wh-pronoun objects are fronted in "Who do you see", but not when part of a conjoined NP. Furthermore, this is not a parochial fact of English, but is true of a number of other languages such as Spanish and German.

(1) *English* *Who$_j$ do you see Mike and $\_\_{}_j$?
Spanish *Quien$_j$ tu ves Miguel y $\_\_{}_j$?
German *'Wem$_j$ siehst du Karl und $\_\_{}_j$?'

The principle is proposed to be a universal property of human language. Conceptions of universals differ as to their strength, and as to what they are statements about.

1.1. Three Concepts of Universals

The strongest position on linguistic universals is set forward in generative grammar. A central tenet of Chomsky's generative grammar is that there are properties true of all human languages. In Chomsky (1965: 27-8), it is stated:

> 'A theory of linguistic structure that aims for explanatory adequacy incorporates an account of linguistic universals, and it attributes tacit knowledge of these universals to the child.'
>
> '... the main task of linguistic theory must be to develop an account of linguistic universals that, on the one hand, will not be falsified by the actual diversity of languages ...'

Under the typical generative interpretation of 'universal', the

study of universals can be the study of the formal properties of a generative grammar for *some* language. As expressed in Chomsky (1965: 28), 'The study of linguistic universals is the study of the properties of any generative grammar for a natural language'. In other words, the study of language universals does not require the investigation of more than one language. Chomsky states further (p. 66, fn 2) that 'Study of a wide range of languages is only one of the ways to evaluate the hypothesis that some formal condition is a linguistic universal'.

Does it make any sense to posit universals based just on English, the language used most widely at least for syntactic research? It does, to some extent, because there are different senses of linguistic universal, and Chomsky's "universals supported by one language" view is to some extent appropriate for a particular type of universal. On the one hand, there are factual properties true of all languages. For example, no language has more than 6 consonantal distinction governed by the state of the larynx; no language forms yes-no questions by exactly reversing order of words. We may term such relatively theory-neutral empirical statements "observational universals".

(2) Observational universals: empirical observations true of all languages
   a. No language has more that 6 manners of consonantal distinction governed by the state of the larynx
   b. No language forms yes-no questions by exactly reversing the order of words
   ('This is an example sentence.' → 'Sentence example an is this?')

Some linguists concentrate on what could be called "pretheoretical data patterns" of this type which are attested across languages, looking for universally true statements.

There is also the abstract architecture of Universal Grammar

(which is itself a theoretical assumption), and this determines to some extent—whether the determination is partial or complete is an open question—the form of the grammar of a specific language. There is ongoing debate over what these principles are, but to give some specific examples, this would include the principles that metrical feet dominate syllables, or that syntactic movement can only move an element from a immediately lower clause. Such universals can be termed "architectural universals". Certain linguists follow Chomsky and focus on these formal properties of the theory of grammar, i.e., architectural universals.

To bring in another major trend on universals, there are two kinds of "observational universal". Some, like the lack of languages with mirror image transformations, are true universals, in that for all languages, they are true, and we can call them "absolute universals". Another type of universal is the so-called universal tendency, exemplified in the research of Joseph Greenberg and his followers. Greenberg, Osgood & Jenkins (1963: 15) claim:

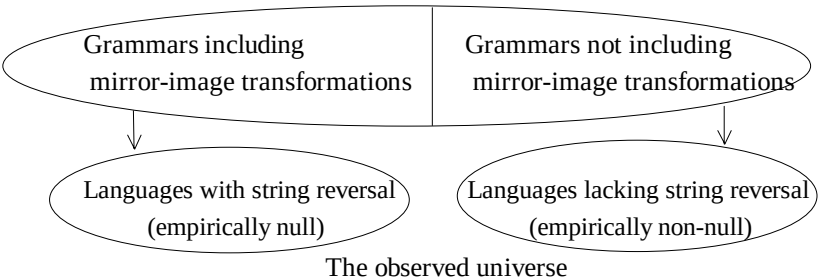
> 'Language universals are by their very nature summary statements about characteristics or tendencies shared by all human speakers. As such they constitute the most general laws of a science of linguistics.'

Note the use of the word 'tendency', rather than 'strict law'. As an example of a tendency, in most languages in the world, word order is either SVO, SOV or VSO. Greenberg (1963) states that the word orders VOS, OSV and OVS either do not occur at all or are extremely rare. Nevertheless, three languages are actually mentioned—Coos, Siuwslaw, and Coeur d'Alene—which have VOS order. The Greenbergian view of universals is thus much broader than the Chomskian one, since the Greenbergian concept of universals includes observational statements which are false for some languages, but which also said to hold true with far greater frequency

than would be expected if the property were determined by chance.

These conceptions of universal can be related, if not entirely unified. The clearest connection is between architectural and observational absolute universals. Architectural universals partition the infinity of mathematically imagineable grammars into one subset, each of which obeys some particular principle, and a complement subset each of which violates that principle. The formal extension of a grammar—the set of linguistic forms that the grammar can describe—is a language, so in partitioning the class of imaginable grammars into those conforming to vs. disobeying a principle, the implicit claim is that the principle predicts two sets of languages in terms of allowing vs. precluding classes of data. Hence the class of imagineable languages which form yes-no questions by reversing word order stands in contrast with the complement class of languages which form yes-no questions in other ways.

(3) The logical universe of architectural principles



The observed universe

The connection between formal theory and observations about languages is made by dint of the fact that the lefthand class of languages would require a special operation in the grammars which describe them, namely a mirror-image transformation. The mirror-image transformation is not included in the universal toolbox that defines most formal theories of grammar. This architectural princi-

ple—that grammars cannot contain mirror-image operations—then *explains* the observational universal that no language actually forms questions in this way.

It is important to bear in mind that a grammar generates a specific language (whether or not such a language is actually observed), but a given, specific language can be generated by a number of distinct grammars. A grammar maps deductively to one language, but a specific language inductively can be the result of a number of grammars. This limits the utility of languages for testing theories of grammar, since an actually observed language may be consistent with (deducible from) very many grammars that are consistent with an assumed metatheory of grammatical rules, and could have been generated by even more grammars, when considering the grammars allowed by all metatheories of grammar.

The relation between absolute observational universal and statistical tendencies of the Greenbergian type is that these are both statements about observations, i.e., they are statements about languages and not about grammars. An absolute universal is one where 100% of human languages have a certain property, and a statistical tendency is one where a significant number, but not 100% have the property.

There is also a connection between architectural and statistical universals. For the moment we ignore the question whether formal theory *should* account for statistical observations; it is an unquestionable fact that the theory of grammar *has* attempted to do so. The essential mechanism for connecting statistical universals and universal architecture is the concept of grammar evaluation, seen as a tool in the process of language acquisition. The idea behind this approach has been that a child learns the simplest grammar possible, and the theory or grammar provides a definition of 'simplicity'. Formal simplicity covers probability, via the special theory of markedness, which allows common processes to be made formally simpler. Some processes, which are said to be 'marked', become more

complex to state formally, and are at a disadvantage in terms of acquisition, and thus they are more likely to be replaced with a process, which is simpler given the theory of formal markedness.[1]

1.2. What Universals are Predicated of

The presumption that there exist linguistic universals has almost taken the status of the null hypothesis in generative linguistic research over the past half century. The actual null hypothesis is that any kind of imaginable coding system could in principle exist as a human language with no restrictions on the frequency of one type of language existing versus any other type. Any discovery to the contrary would therefore be of interest. Despite the explanatory utility of the hypothesis that there are linguistic universals, it is scientifically surprising that this is actually so. The fact that it might be possible to attribute an observation to a special status—being a universal—does not mean that it should be automatically, since that begs the very interesting question of what kinds of facts need to be explained by appeal to universal status. The well-justified existence of one universal does not conceptually license the unlimited addition of statements to the ranks of universal.

---

[1] The topic of markedness will not be pursued here, but a problem in the logic of the theory will be mentioned. An assumed principle of acquisition which relates simplicity to frequency is that a child acquires the simplest grammar consistent with the data. Given two extensionally equivalent rules, one of which is more complex in the formal sense, the simpler rule will be the one learned. This is only effective in cases such as choosing between a rule stated in terms of "voiced sonorants" as opposed to the simpler expression "sonorants", in a language which has only voiced sonorants—i.e., this is just Occam's Razor. The problem is that if the data supports a rule turning /t/ into [s] after a vowel (the marked outcome, according to Chomsky & Halle 1968), then it does not matter at all that a hypothetical rule turning /t/ into [$\underset{\smile}{s}$] would be formally simpler, because selection of the simplest rule is subordinate to selection of a rule that accurately describes the facts. If in a language /t/ becomes [s], the alternative of /t/ becoming [$\underset{\smile}{s}$] is a hypothesis that simply would never be entertained in the process of acquiring a language.

Note too that the null hypothesis is not that all imaginable languages do exist and that they do exist in equal numbers according to language "type", but that they "could" so exist. If the hypothesis were that all imaginable languages actually exist, the hypothesis would be rejected as false since there are infinitely many imagineable languages and finitely many actual languages. Implicitly, the attention of universalist linguistic research has been on identifying general types of imaginable but nonexistent languages, claiming that there are no existing tokens of some type. Many nonexistent human language types are self-evidence, for example languages which employ the release of chemicals as part of the system of signalling, or the use of electromagnetic energy, or acoustic signals above 40,000 Hz—these are modes of communication which cannot be produced and perceived by humans.

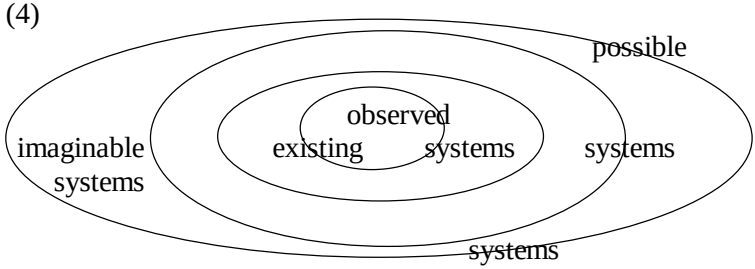
Physical limitations on human languages (e.g., we do not speak in x-rays) are not without interest to the linguist, though this particular fact may be of less interest. Principled limitations on human language are all based on physical fact, be it the electromagnetic fields that humans generate, limits on acoustic frequency resolution due to the physical characteristics of human auditory processing, or brain-structure limits on relations between extracted noun phrases and their traces. Certain restrictions on the types of languages which we claim could exist are very well understood; others, such as those pertaining to the relationship between surface positions of NP's and their thematic roles in sentences are not so well understood. The investigation of linguistic universals is therefore not a search for principled restrictions on language which lack a physical basis. Principled restrictions have a physical basis: the search for universals especially in the context of generative grammar is the search for a particular type of physically-based restriction, namely a restriction having to do with brain structure, in particular those structures that are relate to the human linguistic faculty.

Even ruling out impossible modalities and mental feats requiring

inhuman cognitive abilities, the set of "pre-theoretically plausible" human languages is huge, if not infinite—yet the set of human languages is finite. The standard emphasis on "possible" languages, as opposed to actually attested languages, creates a significant epistemological problem in evaluating the correctness of linguistic research, from the perspective of Popperian views of theory testing. A theory is falsified if it claims that a certain entity exists but the entity does not in fact exist. Any predictive linguistic theory predicts many languages which are not known to exist, so by Popperian criteria these theories are, as far as we know, false.

The relation between observed, existing, possible and imaginable languages is charted in (4).

(4)



Least problematic in this respect are existing but unobserved systems. With diligent linguistic survey work, the set of observed languages could in principle become the same as the set of existing systems. However, linguistic theories usually strive for a higher standard, one of prediction, because we know that new types of languages can some into existence over time. The grammatical system which we refer to as "Modern English" did not exist 1,000 years ago, but it does exist now, and it has different structural properties from its predecessor. Since previously nonexistent grammatical systems are constantly evolving, it would be quite shortsighted to develop a theory which only explains the nature of currently existing lan-

guages, thus linguists focus on the idealisation "possible language". An example might be a language which is exactly like English, except that the verb of main clauses is placed at the end of the clause—such a language does not seem to exist at the moment, but it is likely that this is a possible language. The implicit assumption behind the concept possible language is that although some particular possible language may not exist at the moment, given enough time that language will eventually come into existence by random developments in existing systems. The third category of unobserved languages is the set of languages which we say not only do not exist, but also cannot ever exist (barring evolutionary changes in brain structure that necessitate changes in the concept possible language).

Since we only have a set of actually observed languages to work from, the question is whether there is a reasonable possibility of fleshing out the concept possible language which includes not just observed languages, but also includes languages which do not even exist. The study of linguistic universals provides the inductive foundation for a predictive theory which actually states what a "possible language" is.

2. The Statistical Foundation of Universals

Before trying to list potential universals of human language, we need a valid basis for identifying universals. Since the logically prior question in this matter is evaluating evidence in the search for universals, I start with statistical tendencies, since predictive science is based on inductive generalizations from observation to the broader class of actual existents. Inductive generalizations are grounded in probability, and investigation of statistical tendencies makes these central issues of probability most explicit. Even supposedly non-statistical absolutist views of universals have a soft statistical underbelly when it comes to the evaluation of unrefuted hy-

potheses, as we will discuss in the third part of this section.

2.1. Improbable Events

As Greenberg has pointed out, with greater than chance frequency, human languages select one of the orders VSO, SVO or SOV as their basic word order. The crux of the issue is summed up by Comrie (1981: 19-20) who says:

(5) '... the disparity between the number of languages violating the universal (probably less than 1 per cent of the world's languages) and those that conform to it is massive. To say that the universal has no validity because there are counter-examples to it, and leave the discussion at that, would be to abrogate one's responsibility as a linguist to deal with significant patterns in language.'

Payne (1997: 76) similarly states (here A is roughly "Subject" and P is roughly "Object") that "The tendency for A to precede P in basic, pragmatically neutral clauses is so overwhelming that it it extremely unlikely that it could have arisen by chance". Contained herein are, apparently, testable empirical claims as well as an important philosophical point about the goals of science: is it true that only 1% of the languages of the world violate the basic word order law, and why would that fact be of interest, if it is true?

The null hypothesis is "there are no regularities; data is distributed randomly". Scientific research seeks to reject the null hypothesis by finding the actual regularities that exist in the universe, as long as there is objective justification for claiming that a regularity exists. If word order were randomly distributed, we 'expect', in an idealised mathematical sense, that in a sample of 300 randomly selected languages, we will find 50 languages with each of the 6 logically possible orderings of subject, verb and object. The condition

that the languages be randomly selected is important. If we select 300 Bantu languages, we will find no word order other than SVO, due to a historico-genetically explained fact about these languages.

Were we to actually observe such a "50 cases of each type" distribution of word orders, we would have no warrant to draw any special conclusions about word orders. In this case, the observed distribution (1/6 = 16.66% of the total sample for each case) exactly matches the theoretically expected frequency that would result if word order were random, so the probability that the observed pattern is identical to a purely random pattern is 1: it is certain that the pattern is random. Nor for that matter could we legitimately assume that there is anything special about certain word orders if we observed the orders SVO, SOV and VSO in 51 languages apiece, and observed the orders VOS, OSV and OVS in 49 languages apiece: the observed frequencies (17% and 16.3% of the total sample) is so close to the predicted frequency—the observed-to-expected ratios are 1.02 and .98, which is nearly identical to the perfect random distribution—so we must conclude that this small divergence from the ideal random distribution is of no importance.

The probability of finding 99 languages with each of the three unmarked orders and 1 language with each of the three marked orders, a clearly non-random distribution, can be computed by statistical procedures, and this reveals that there is only 1 chance in 3×10^{60} that this distribution is by chance—this is an extraordinarily improbable outcome. If presented with such a pattern of word orders, we would have to conclude that word order is not randomly distributed. This massive disparity in the distribution can be expressed as a statistical significance value, which provides an objective basis for rejecting the hypothesis that the word orders are not distributed at random. There are various conventions for relating those probabilities to epistemic states such as "certainty", for instance the conventional standard in social science that a distribution which stands less than a 1 in 20 chance of being due to random chance is not in fact to

be considered random; or one could adopt the more stringent standard of less than 1 chance in 100. Certainly a probability of less than 1 chance in 3×10^{60} qualifies for "certainty".

In truth, this 1% figure conjectured by Comrie for the distribution of OVS, OSV and VOS orders was a hunch based on the author's experience with a broad range of languages, and was not the result of a careful search and statistical tabulation. Indeed, Payne (1997:76) indicates that 70% of languages are SVO or SOV, and 15% are VSO, suggesting that the 1% estimate may not be totally accurate. However, we have no reason to think that an exhaustive search would completely overturn the generalization that the observed rareness of the rare orders is not due to chance, even if the exact degree of non-randomness of the distribution is subject to refinement. In principle, an actual observation-based figure could be computed, by determining word order in a large number of languages, calculating the observed frequency of each order, and computing the probability of the particular distribution. But even if this had been done, there is a very good reason to be skeptical about any such results.

2.2. Nonrandom Sampling

A fundamental problem with statistical universals is that their validity depends crucially on something that simply does not exist, namely a database drawn from a collection of languages where all existing human languages have an equal chance of being represented. A fundamental principle behind the concept of sampling which statistical extrapolation depends on is that probability inferences from a sample to a population (e.g., "the set of existing human languages") are valid only if all members of the population have an equal chance of being represented in the sample.

It is invalid to look at structural properties in 6 Romance languages and extrapolate from statistical tendencies observed in Ro-

mance to general probability that the properties will be found in all human languages. While such blatant violation of the "equal chance of representation" principle would never be consciously countenanced, that assumption is nevertheless systematically violated in statistical typological research on languages. Certain languages *do* have a far greater than chance probability of appearing in any sample of languages used for statistical estimations; not all languages have an equal chance of being selected.

Greenberg in his famous word-order universals paper was very probably aware that his sample was highly biased, and did not attempt to make inferences based on formal computation of probability of appearance at random—e.g., he never computed actual numeric probabilities that the patterns found in his survey could have arisen by chance. However, Greenberg's universals frequently make rhetorical appeal to the presumption of such statistically-supported inferences, with such phrasing as "overwhelmingly greater than chance frequency" being part of the statement of universals 4, 9, 17, 18 and 22.

Consider the 30 languages which Greenberg used.

(6) Indo-European languages:	Greek, Italian, Norwegian, Serbian, Welsh, Hindi
Niger-Congo languages:	Fula, Swahili, Yoruba
Austronesian:	Malay, Maori
Nilo-Saharan:	Masai, Nubian, Songhai
Afro-Asiatic:	Berber, Hebrew
Others:	Basque, Burmese, Burushaski, Chibcha, Finnish, Guarani, Kannada, Japanese, Loritja, Maya, Quechua, Thai, Turkish, Zapotec

Included in that list were six Indo-European language: Indo-European languages constitute 20% of his database. But the existing

Indo-European languages constitute only 6.6% of languages of the world. [2] Furthermore, the non-Indo-Iranian languages of Indo-European make up only 2% of the world's languages, but they are 20% of the languages in Greenberg's survey, a ten-fold over-representation of Indo-European languages.

On the other hand, the roughly 1,400 Niger-Congo languages constitute 20% of human languages, but the three Niger-Congo languages in that sample represent just 10% of the database. Austronesian languages constitutes 18% of languages of the world but a mere 6.5% of the database. Interestingly, the sample is most representative, statistically speaking, with Afroasiatic languages, which makes up 5.5% of human languages and 6.6% of the sample. Ironically, Nilo-Saharan languages makes of less than 3% of the languages of the world, but 10% of the languages in the database, so an obscure language group is being over-represented. The Trans-New-Guinean languages which make up about 8% of languages are not represented at all, yet the sample, were it representative, should include two such languages. In this sense, the database is *not* representative of the world's languages.

One way to address this problem is to demand that all statistical inferences about language be based on a sufficiently large and truly random sample of the languages of the world, which would make such an uneven distribution virtually impossible. This ideal proposal is guaranteed to fail in actual practise, because documentation simply does not exist for all languages. The problem with Greenberg's sample was simply that no random sample can be practicable. This over-representation of Indo-European and under-representation of Niger-Congo was purely a function of what information is available on languages.

To get an idea how serious the problem of under-documentation

[2] Information about numbers of languages in the world and numbers of speakers has been taken from the Ethnologue web pages (http://www.ethnologue.com) and its predecessor, the SIL web pages.

of languages is for forming statistical generalizations about human languages, we will take 3 random samples of 20 languages each, out of the pool of human languages (derived from the list of languages on the Ethnologue website). If enough languages are well documented, it may not matter that some languages have inadequate documentation. The procedure was to pose a simple factual question about language structure, and determine the answer to the question using published materials available in a major research library (the one at Ohio State University). For the first group, the question was how often languages have only one series of obstruents in terms of laryngeal properties (as Hawaiian does), or more than one (as English and Korean do).

(7) Sample 1: Tangoa, Korana, Kanamari, Saban, Zauzou, Wushi, Worora, Malaynon, Judeo-Tat, Batak Karo, Turaka, Cofan, Kasiguranin, Wangganguru, Motu, Teke, Kati, Wabo, Zoque, Babine

Languages with available material: Korana, Worora, Judeo-Tat, Batak (only the Toba dialect), Wangganguru, Motu, Kati, Zoque, Babine

It was possible to locate relevant information on only 9 of languages, less than half of the target group. Any statements about this aspect of human language based on this sample would be strongly biased in favor of the subset of languages that have benefited from linguistic description: not all languages have an equal chance of being represented in the sample.

In the second set, the question was what percentage of languages have only prefixes versus only suffixes versus both. This question is harder to answer, because it requires a deeper level of linguistic description than the previous question.

(8) Sample 2: Ndut, Yaqui, Lala-roba, Woi, Atikamekw, Campalagian, Zari, Djongor Bourmataguil, Kyaka, Tondano, Nomatsiguenga, Waci-Gbe, Hinihon, Mape, Tugutil, Otuho, Khuen, Buwal, Nyong, Tandia

Languages with available material: Yaqui, Kyaka, Tondano, Nomatsiguenga, Waci-Gbe

Materials of any sort were only located for 5 out of the 20 languages, which is such a low success rate that there is no point in counting numbers of languages.

The third question was, how many languages allow wh-movement out of subordinate clauses where the lower subject has been raised into object position in the matrix clause—where two NP's raise out of the lower clause—as exemplified by the English sentence "Who does Bob believe Tom to have explained the solution to?". Descriptive materials were found for only for 8 out of this third set of languages. Disregarding the problem that again a representative random sample of the required size cannot be found, those materials were searched in depth, in the hopes of determining how many out of that (unrepresentative) set do allow this construction. As it turned out, none of the source materials were rich enough that a determination could be made for even one of these languages.

(9) Sample 3: Serili, Mingrelian, Yahadian, Anuak, Noon, Mundu, Malagasy, Melokwo, Temiar, Mazahua, Laba, Amdang, Grangali, Miltu, Kwaya, Ijo, Kuuku-yau, Saraiki, Tswapong, Malaryan

Languages with some available material: Mingrelian, Anuak, Mundu, Malagasy, Temiar, Mazahua, Ijo, Tswapong

Languages with **sufficient** material: ∅

The crux of the problem is that documentation of human languages is highly asymmetrical, with respect to certain politico-cultural features. We have a pretty good knowledge of the structure of most European languages (at least the main national dialects), the Semitic languages (at least Arabic and Hebrew), and various classical literary languages such as Chinese, Japanese and Korean. We know comparatively little about the structure of the languages of Africa, and next to nothing about the languages of New Guinea. Furthermore, the quality of documentation is quite uneven across languages.

A well documented language is one where the facts are clearly stated and exemplified in such a way that relatively few empirical questions cannot be resolved on the basis of published materials. Such languages are very useful in the testing of linguistic hypotheses. A poorly documented language is, analogously, one where many questions cannot be resolved from the published sources, and such languages are of less use (perhaps no use, depending on the quality of the information) in hypothesis testing.

English has a good claim on the title "best-documented language". There are numerous tomes on English structure, and no language has received more attention in syntax than English. This is not to say that there are no remaining empirical generalisation lurking out there unobserved: but in comparison to any other language, we know a huge amount about English. At the other extreme of the documentation scale—a well-populated extreme—is the language Ngindo, a Bantu language spoken in Southeastern Tanzania. There is not a shred of documentation on its structure in the published literature; no grammar, no article, no word-list, no Bible translation. For the moment, the scientific community has no idea about the answers to the simplest questions about that language. There are very many languages which remain completely undescribed or vastly under-

described, out of the 6,800 or so languages spoken in the world: perhaps 2/3 of the languages of the world fall into this category.

The documentation asymmetry—the fact that some languages have much more descriptive material than others—is not evenly and randomly distributed throughout the world, in terms of numbers of speakers or geographical location. 54% of the world's population speaks one of only ten languages listed in (10).

(10) Asymetries in language distribution
Top 10 languages: Chinese, Spanish, English, Arabic, Bengali, Hindi, Russian, Portuguese, Japanese, German

Seven of these world's largest languages are members of the Indo-European language family. 95% of the world's population speaks a mere 5% of the languages in existence, an elite group of languages which have at least a million speakers. Put differently, 95% of the world's linguistic diversity is in the hands of a mere 5% of human population. 82% of the world's languages have under 100,000 speakers, and half of the world's languages have fewer that 6,000 speakers. The overwhelming majority of this last group of languages also fall into the undescribed category. Somewhat-well described languages with very small populations such as Amele, Arbore, Klamath, Lakota, Menomini, Miwok, Woleaian and Yurok are unusual. More commonly, very small languages such as Degexitan, Dilling, Hu, Kaningi, Mangole, Merlav, Mlomp, Ndasa, Vilela, Yos, and Yukpa are undocumented.

This diversity is uneven over the world. Around 2,000 languages are spoken in Africa, which is a bit under 30% of the world's languages. The uncontested winner in the diversity contest for one country is Papua New Guinea, with around 820 language spoken by a population of about 5 million, in an area that is only a little bit bigger than the state of California. This is followed by Indonesia

which has 726 languages. Other countries with large numbers of languages include Nigeria with 500 languages, India with 387 and Mexico with 288.

If the rate and quality of linguistic documentation were consistent throughout the world, it would not matter that some languages are better documented than others. But in fact, while the languages of Africa and the languages of the Pacific contitute the majority of the languages in the world, these are exactly the most poorly documented languages, because these are not major literary languages with long literary and academic traditions, and thus do not have the solid and slowly accumulating descriptive basis that major languages such as Chinese, Spanish, English, Arabic, Russian, Portuguese, Japanese, and German (and Hindi, to some extent—Bengali is the relatively rare case of a less-documented language with huge numbers of speakers).

To summarize, statistical inferences about the nature of human languages based on frequency of occurrence of properties in observed languages are valid only if based on a representative sample from the population of human languages. Because of significant skewing in the documentation, there can be no representative sample. Entire linguistic areas and language families are excluded from participating in random tests of many linguistic hypotheses.

2.3. An Argument against Unbiased Random Sampling

It might seem that this problem could be addressed simply by an increase in descriptive work on under-studied languages. Based on the failure rates observed above in finding linguistic documentation, we can estimate that approximately 1/3 of the languages of the world have been sampled to a small degree as a result of a few centuries of previous research on languages, which gives one an idea of the size of the task ahead in filling in the remaining gaps. While such a move will be of great value to understanding the diversity of

human language, there is another problem of principle for the program of discovering statistically significant tendencies, and that is the problem of genetic asymmetries.

Since languages are in the heads of individual speakers, then if one has a cognitive interest in the nature of language, why not take a random sample of humans, and extrapolate from properties of the grammars in their heads to the properties of human language—the set of all grammars in all the heads of humans? The problem is that we would discover that, with greater that chance frequency, the properties of human language are the properties of Chinese, because Chinese speakers constitute a large portion of the humans in the world. Nobody would advocate making inferences about the nature of human language in a way that allows massive leveraging by this extraneous socio-historical fact about population. For this reason, linguists focus on languages and not speakers, and treat all languages as being equal, each deserving of a single vote. But counting languages is actually like counting speakers. Linguists implicitly seek generalisations that are free from the influence of historical nonlinguistic accidents. Just as a proper statistical view of language shouldn't yield a theory of universal properties excessively leveraged by the number of speakers of Chinese in the world, we also do not want our theory of what counts as "common" in human language to be distorted when a particular genetic group of languages predominates in our sample of languages.

A widely-held assumption regarding language universals is that they reveal something about human nature, either about human cognitive abilities, or something about an innate language faculty. Taking a large sample of languages and looking for significant patterns seems like a reasonable way to get this information which has deeper significance. For such an inference to be valid, we must be assured that no hidden factors reduce supposed universal properties to the status of a linguistically insignificant coincidence. Genetic factors are a very significant hidden factor affecting the validity of

hypotheses based on language samples.

Here is an example of how genetic biases in samples can lead to incorrect conclusions. The frequency of rules of vowel harmony which affect the features [back] and [round], versus harmony in vowel height, may be of theoretical importantance because a theory of vowel features proposed in Archangeli (1985) predicts that there should be significantly more languages with back-round harmony than with height harmony. That article proposes a particular geometry for vowel features which is justified by the fact that it accounts for this asymmetry in occurrence of the two kinds of harmony: (p. 369) "If we assume that altering structure prescribed by Universal Grammar is costly, then we account for the rarity of rules spreading [high] independently of [back, round]".

The argument is predicated on the assumption that Height harmony is in fact less common that Back-Round harmony. To check that claim, I performed a count of all languages which I could identify with either kind of harmony. The result was the discovery of 24 languages with Back-Round harmony, but 86 languages with Height harmony. Clearly, the foundational assumption was wrong. Indeed, based on this new information about frequency of occurrence, an organisation of phonological features exactly the opposite of that proposed by Archangeli would be better supported, at least if frequency of occurrence is related to formal linguistic structure, as Archangeli assumes via appeal to the theory of markedness and cost-assignment.

The explanation for this contradiction is that almost all of the examples of Back-Round harmony are from Uralic or Altaic languages, and almost all the languages with Height harmony are Bantu languages. Either due to genetic inheritance or areal diffusion, most Altaic languages and many Uralic ones have Back or Round harmony; thus these languages are predisposed to having that kind of harmony for a reason that has nothing to do with the theory of grammar. Similarly, there was probably a height harmony rule in Proto-Bantu (or an early branch in Bantu), and this rule was faith-

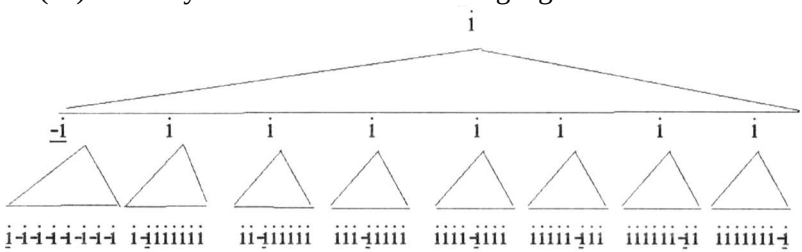
fully inherited by most Bantu languages. The numeric asymmetry exists because there are ten times more Bantu languages than Uralic and Altaic languages. In this case, the wrong conclusion was reached about what is more common because a fairly high percentage of the Uralic and Altaic languages have or had been subject to in-depth grammatical description, compared to Bantu languages.

If the goal of statistical interpolation based on language sampling is to provide knowledge of the natural propensities of human language, free from the influence of linguistically irrelevant coincidences, then a random sampling of human languages would always over-represent Bantu languages, insofar as Bantu languages have a ten-fold better chance of appearing in any sample than do Uralic or Altaic languages: Bantu languages make up about 7% of the world's languages. Given that the language "Proto-Bantu" had a rule of vowel height harmony, we can predict that most descendants of Proto-Bantu will also have height harmony (it should be born in mind that Bantu is a relatively homogenous language family with a time depth comparable to Romance). The predominance of height harmony in human language—as would inevitably be revealed by a random sample of human language—is therefore just as likely to be due to this socio-cultural historical accident, as it is to be due to a fundamental structural property of human language.

The reason why we need to be concerned with over-abundant language families is that languages are fundamentally conservative across time. If a language has a given property, it is very likely that the property will be preserved in all languages descending from it. Consider an ideal situation graphed in (11). The chance that a linguistic property will change over a short interval of time is very small. Let us say that the inherent probability of change in the space of one generation is 1 in 8—surely this number is way too high, but it will do for the purposes of making calculations. Starting with one language, a proto-language, if this language gives rise to 8 daughter languages, and changes in a property *i* occur at the expected rate,

then 7 of the 8 daughters will retain the original property *i* and 1 in 8 of these daughters will have the changed property, *-i*. Each offspring in turn generates descendants, and the properties of the parent are correctly replicated in the offspring seven times out of eight. We can observe the distribution of the property in the third generation descendants—a set of hypothetical actually observed languages — given this perfect distribution of descendants, the original property is retained 78% of the time (a daughter which undergoes a change from its mother language is underlined).

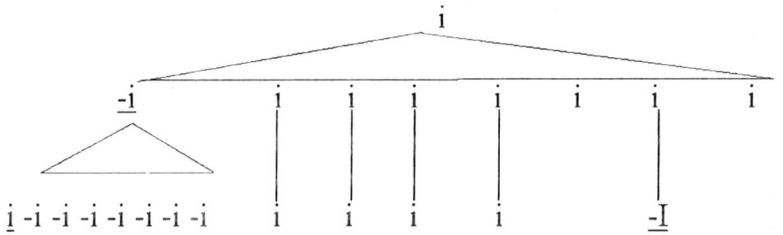
(11) Perfectly Even Distribution of Languages



Given this uneven distribution of the property *i* in observed languages, we are also warranted to make the inductive inference that the original language indeed had the property *i* (and not the property *-i*).

Suppose, however, that only one second-generation descendant is highly prolific—some daughter languages die out leaving no modern descendants, and most others leave only a single descendant. By chance the highly prolific daughter language happens to be the one exhibiting the change in property *i*, as in (12). The consequence of this is that in the third generation, the original property *i* appears unchanged only 38% of the time, and well over half of the language attest the contrary property *-i*.

(12) Uneven Distribution of Languages



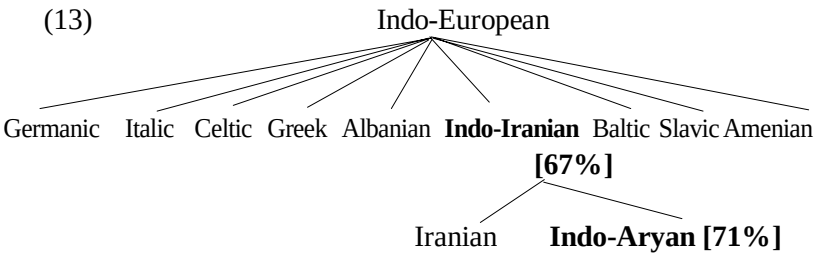
By the same type of inductive reasoning from properties of the observed set to properties of the general class, we would conclude that the original system most probably had the property *-i*, since the property *-i* is observed 62% of the time. In this (hypothetical) case we know that the conclusion is wrong, and we know the confounding factor that explains why an original system with the property *i* appears to have changed to the opposite property so often: some linguistic subgroups are more prolific than others.

It is quite reasonable to assume that humans evolved linguistic capacity once, and that all human languages are ultimately related to each other, so the statistical study of the frequency of properties in human language does face the problem (12), in terms of our ability to use observed frequency as a basis for forming a valid inductive hypothesis regarding causality. One possibility that cannot be rejected out of hand is that some properties found in many human languages are true of these languages simply because the proto-language of all languages had those same properties; they may reflect accidental linguistic conservatism rather than deep cognitive properties. Since no causal law require linguistic properties to change after a certain number of millenia, statistically frequent properties of human languages could be as much an accident of history as the predominence of height harmony is in Bantu languages. Thus our estimates of what properties are frequent in human language is highly leveraged by the basically conservative nature of

languages over time and the uneven distribution of related languages in the set of existing languages.

The possibility of an improper leveraging due to asymmetries in the number of languages according to genetic criteria is not just a theoretical possibility: such asymmetries are common. Indo-European is just one of about 100 linguistic phyla in the world. Within Indo-European, 67% of the languages belong to the Indo-Iranian family, which is just one of the 9 (surviving) branches of Indo-European. Within Indo-Iranian, the Indo-Aryan languages, more or less those descended from Sanskrit, constitute 71% of the Indo-Iranian languages.

(13)



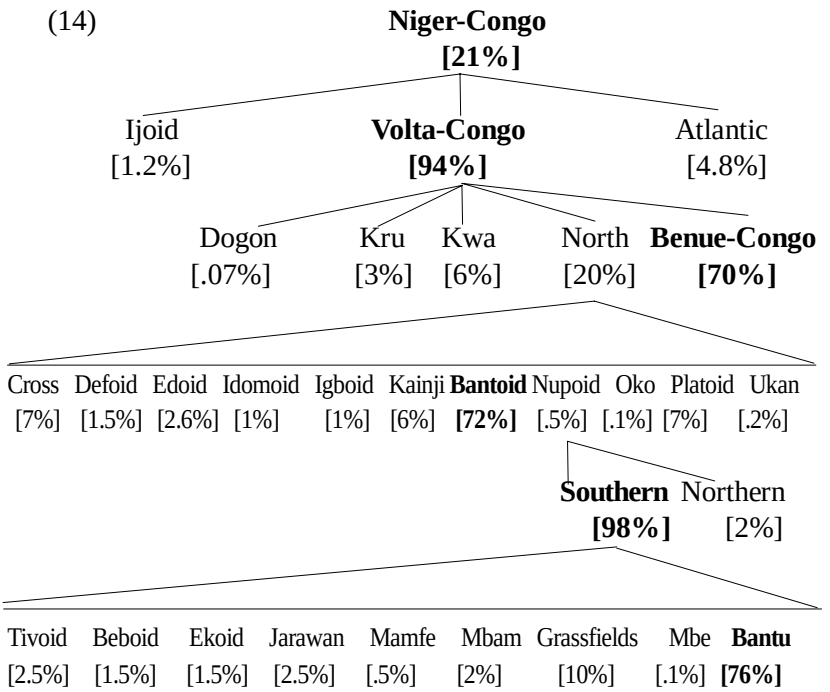
If we were to sample Indo-European languages at random, the most common properties observed would be those found in the Indo-Aryan languages.

The biggest skewing in terms of numerical distribution is the Niger-Congo phylum, whose roughly 1400 languages account for 21% of the languages of the world. Within Niger-Congo, the Volta-Congo group is only one of 3 subgroups, but it contains 90% of the Niger-Congo languages. Thus historical changes that affected the Volta-Congo proto-language have a very good chance of being passed on to any descendant. Phrasing the problem differently, because there are so many Volta-Congo languages, statistical reasoning would lead us to believe that what happens to be true of Volta-Congo is very probably true of Niger-Congo, or even true of all lan-

guages. Volta-Congo is composed of 5 historical sisters, but the Benue-Congo branch accounts for 70% of those languages. The Bantoid languages define 1 of 11 subgroups and make up 72% of the languages: most of those languages are Bantu.

An analogous problem arises with the Austronesian language phylum, whose 1,200 languages make up about 18% of the languages of the world; or, the Trans-New Guinea phyllum which has 540 languages and makes up about 8% of the world's languages.

(14)



The paradox of language sampling is that on the one hand, an unbiased random sample is required for an inference based on observation of a sample to validly extend to the entire population of human languages; but a truly random sample obscures significant

asymmetries in historical causation, thus leading to invalid inferences about human language which are based on excess influence of certain language families.

Dryer 1992 presents a counting methodology which lessens some of these problems. His methodology for detecting word order correlations involves counting language genera with particular properties, in a set of 252 genera. The method looks at e.g. whether VO order and prepositions exists at all in Bantoid—and various other genera—so it does not matter whether that pattern is broadly attested in Bantu (it is) or marginally attested. Potentially, any genus can have as many "votes" as there are possible patterns being investigated. Total genera attesting each possible pattern are presented, with subtotals for six broad areal groups such as "Eurasia" or "Southeast Asia and Oceania". This tells us that there are 16 genera in Africa with VO ordering and prepositions, but only 4 African genera with VO order and postpositions. This methodology might in principle have the advantage that only very robust correlations yield strongly asymmetrical distributions across all language areas, as is the case with the orders OV&postposition ~ VO&preposition. The method thus says nothing about how common a property is in the languages of the world.

There are two major flaws in the method. First, a genus is taken to be a maximal language group reconstructed to a time depth of no more than 4000 years, by which criterion Baltic, Slavic, Italic and Celtic are separate genera (despite being daughters in Indo-European). As Dryer notes, information about time depths is quite meager and involves a lot of guesswork. It is also subject to significant disparities in the amount of historical reconstruction which has been done on languages in various parts of the world, especially at such a time depth. Second, undersampling of languages within genera means that the appearance of uniformity (a false positive with respect to the question whether certain correlations are robust) can result when a genus is actually non-uniform. Many genera in the da-

tabase are represented by single languages: for example 27 of the 47 African genera and 42 of the 70 North American genera are represented by one language. If all four word-order possibilities are actually found in some genus but only one language is sampled, it is impossible to tell that the supposed correlation actually fails completely in the genus, as opposed to being rigidly observed. The only language representing Surmic, one of the African genera, is Didinga; but Surmic is a typologically quite diverse family for matters of basic word order—see Dimmendaal 1998. Nonetheless, this methodology moves—correctly, in my view—away from indiscriminate frequency measures.

3. Absolute Universals vs. Statistics: The Conduct of Theoretical Linguistics

The preceding section argues that numerical extrapolations based on the properties of samples of human languages are fundamentally flawed, because they are not, and from a practical point of view cannot be, based on random samples of human languages, and yet they must be, in order for those numbers to be validly generalized to all human languages. At the same time, genetic (and areal) relation is a major hidden variable which cannot be eliminated or ignored, but which also cannot be corrected for, at least given current fragmentary knowledge of the genetic relations of all human languages. The conclusion I draw from this is, very simply, that numeric techniques with their implication of certainty are unsuited as a method for establishing fundamental regularities about human language. Statements of the type “32% of the languages surveyed have front round vowels” can only be legitimately taken to be a summary statement about the specific languages considered, and not a statement about human language.

3.1. Implicit counting

The problem is actually somewhat worse for the quasi-statistical approach used by many theoreticians, who often use implications about frequency and probability to support particular theoretical claims, without providing actual counts, calculations, or establishing particular significance values. The literature is rife with statements about what is common, and the concept of frequency of occurrence is widely invoked, but these statements are usually presented without giving a count of languages. We noted that Comrie (1981) had an opinion about word order in "probably less than 1 per cent of the world's language". Similarly, Archangeli (1985) expressed the view that certain kinds of vowel harmony are rare. Sagey (1986) argues that phonological processes which occur with high frequency in languages should have certain formal properties.

(15) 'Another requirement on the theory is that the relative simplicity of describing in the representation each process or form that occurs should reflect its relative naturalness, in the sense of its frequency of occurrence in the languages of the world.' (1986:1)

'… with the feature matrix, rules affecting a group of features are more complex and are predicted to be less common than rules affecting a single feature; but in reality, such rules are at least as common as rules affecting only one feature.' (1986:7)

The implication here is that we have reasonably good information about the frequency of occurrence of rules of different types, according to how they are formulated.

Clements also makes claims which are statistical at heart:

(16) ‘se two features ... tend to be interdependent in most classes of sounds in most languages.’ (1985:230)

‘... but we know that rules of the latter type are extremely rare, if not unprecedented.’ (1985: 237)

It is taken for granted that we already have established these frequencies; but in reality, counts are systematically lacking, and for the reasons I have discussed would be of questionable validity as statements about the nature of human language.

3.2. Refutability and Absolute Universals

The biggest problem with statistical universal claims is they they have a dubious empirical status. An absolute universal claim about language makes an unqualified non-existence claim: it states that no language will have the relevant excluded property. For example, the universal feature constituency proposed in Clements & Hume (1995) precludes the existence of a even one single rule which operates simultaneously on the features [labial] and [voice], and the discovery of one remarkably rare language which has such a rule would definitively refute the theory. The significant advantage of an absolute universal, from the epistemological point of view, is that a single counterexample has absolute probative value.

The refutation of a frequency-based universal is much more difficult even under ideal circumstances, which makes statistical universals universals intrinsically less interesting as scientific claims. A statement such as “languages that change the voicing of consonants after nasals usually make voiceless consonants voiced” is not refuted by the fact that the Sotho languages have rules of post-nasal devoicing. The claim is that postnasal devoicing is unusual, not impossible. In order to overturn this implicitly statistical claim, one would have to show that there are so many counterexamples that the

frequency of postnasal voicing does not even rise to the level of "usual" behavior, and that from a statistical point of view the probability of post-nasal voicing is not significantly different from the probability of post-nasal devoicing. As we have seen, because of documentation asymmetries, statistical inferences from a sample of languages to the even less ambitious class "existing languages" are invalid since something on the order of 2/3 of human languages have no chance whatsoever of being included in a sample, given that they are undescribed. Furthermore, genetic relation and over-representation of certain types of languages is a variable which is not controlled for, and if these factors are not controlled for, conclusions based on skewed samples are not valid.

To avoid the problems inherent in statistical approaches to universals, the obvious solution is to reject statistical universals, and focus instead on clear-cut absolute universals. Absolute universals would seem to rest on scientific bedrock because they can be cleanly refuted and rejected based on one (suitably argued) counterexample. Unfortunately the aforementioned problems of under-documentation and information skewing still plague an absolutist view of universals.

The problem for absolute universals arises not in deciding if a statement is false, but in deciding if it is true. While reality itself only presents "things that are"—corresponding to true statements—and "things that are not"—corresponding to false statements— there is an extensive epistemological middle ground, populated by statements which we lack compelling evidence to relegate to the "true" pile or the "false" pile. While propositions (as statements about reality) are two-valued—true or false—knowledge is (at minimum) three-valued—"yes", "no" and "I don't know". Just because a theory hasn't yet been refuted, we can't conclude that the theory is true—there needs to be some separate justification for claiming that the theory is true.

Empirical claims often depend on real-world facts which may not always be directly observed. For instance, a statement about

properties of "all human languages", even if only predicated of *existing* human languages, is not definitively true on the basis of observation, since not all existing human languages have been observed. Universal claims about language are inductive projections from an observed set of instances to the whole class of human languages. When a scientist evaluates a statement involving unobserved events, they implicitly consider the probability that were an observation actually made, the statement will remain true after the observation.

A linguist who advances the claim that no language has both central and back unrounded vowels is implicitly saying that they have surveyed an appropriate number of appropriate languages (or know of such a survey), and find no language with both central and back unrounded vowels. This forms the justification for advancing the universal statement that no language has both kinds of vowels. Put this in contrast to a hypothetical linguist who wakes up one morning with a "hunch" that no languages with such a vowel contrast. The latter linguist does not have sufficient empirical warrant to advance the statement as a scientific claim.

We thus have the reasonable expectation that in making such an inductive generalization, the researcher has actually looked at more than three languages if they make a universal non-existence statement about back and central vowels; if this expectation is not satisfied, we would rightly feel swindled. The probability of a coin landing heads-up three times when tossed three times is 1 in 8, which by usual standards of sufficient improbability is not a highly improbable event, and a claim that the coin is weighted against tails would not be justified given such a meager factual basis. Three coin losses or three languages is not a good enough justification for a predictive claim. Even when "official" statistical procedures involving numeric computation of probabilities are eschewed, the essential problem remains that when we evaluate statements about languages, we presume that the conclusion is warranted by having been tested against

a suitably large unbiased database where there are no hidden variables that turn out to be important explanatory factors.

3.3. The Diversity of the Database in Grammatical Theory

Generative grammatical theories typically make absolute architectural universal claims, such as statements about the types of metrical feet that can exist, or the universal arrangement of distinctive features into higher-level constituents such as proposed in Clements & Hume 1995, or The Minimal Link Condition in syntactic theory. An important question arises regarding the empirical foundation of these and other proposed architectural universals, in light of the issues raised in the preceding sections, is whether they are based on a sufficient empirical foundation. For instance, if linguistic theories are only systematically tested against a half-dozen closely related languages, one might well suspect that such universals are not universally valid. Each universal needs to be tested on its own; some universals are well supported and others are not well supported. My interest in this section is to ask the question of theoretical generative linguistics, in general.

To get an initial estimate of the nature of actual linguistic diversity in theoretical linguistic research, I undertook a survey of articles in the journal *Natural Language and Linguistic Theory* from 1989 to 1998, dividing articles into phonology versus syntax-semantics. The 29 phonology articles make up less than a fifth of the database, and the 128 syntax articles make up the remainder. The reason for making a division between syntax and phonology is that experience with syntax and phonology tells us that there are differences between these subdomains in terms of language coverage. Syntax articles tend to focus on a few languages—the major languages of Europe, especially English—but in phonology, there is less influence in the literature from a very small set of languages.

For each article, examples in the article were tabulated in terms

of which language they derive from. In the phonology articles, 117 languages provide data, whereas in syntax articles, 101 languages are cited; remember too that phonology articles appear at around 1/5 the rate of syntax articles. The number of languages invoked in each phonology articles averages 5, but the average number of languages cited in syntax articles is 2.5. So in terms of sheer numbers, phonology wins the prize for crosslinguistic breadth.

A way to quantify the general influence of a specific language on the literature (and thus get an estimate of the "leverage" which a particular language has on linguistic theory) is to compare the frequency with which the language is cited in articles. The widely-cited languages in syntax articles, appearing in 8 articles or more in ascending order of frequency, are those in (17), with English at the top, which alone figures into 75% of the articles. On the other hand, in phonology articles, the most widely cited language is Polish which appears in only 4 articles: no language figures into at least 8 articles. English hardly figures into *any* articles.

(17) a. Most-often cited languages: Syntax
Japanese, Modern Hebrew, Icelandic, Spanish, Italian, Dutch, German, Chinese, French, English
b. Most-often cited languages: Phonology
Japanese, Spanish, Yoruba, Polish

Another measure of data diversity in the literature is the proportion of the data coming from a given language to the overall pool of examples. In syntax, 20% of the data come from English, followed by French which accounts for 8%. The languages which make up most of the syntax data are in (18a), and these languages as a group account for 2/3 of the total set of examples in syntax articles. Of the 13 languages figuring prominently into syntactic articles, 10 are Indo-European, and 9 are European. In the phonology articles, there are the 19 languages in (18b) whose examples together comprise 2/3

of the total phonology data. Of the better-discussed language in phonology articles, 5 are Indo-European, 2 are Semitic, 2 are Niger-Congo, 2 are Austronesian, and the other 8 are unrelated.

(18) a. Languages accounting for most of the data: Syntax
Hindi, Japanese, Russian, Icelandic, Polish, Breton, German, Irish, Dutch, Modern Hebrew, Chinese, French (8%), English (20%)

b. Phonology
Latin, Asheninca Campa, Slovak, Temiar, Nawuri, Ponapean, Shanghainese, Shona, Arabic, Kashaya, Columbian Salish, Macedonian, Bengali, Indonesian, Yawelmani, Mohawk, Mongolian, Polish, Maltese

At least in syntax, there is noticeably less diversity in the data, with a strong bias in favor of European languages.

There is another side to this coin. Ideally, a linguistic theory is supported not just by good breadth of analysis, but also by good depth. By that latter standard, the tables are turned. Let us say that the depth of coverage for a language is good in an article if data from the language accounts for a high percentage of the data in the article. Using percentage of data from a single language in a single article as an indicator of depth of analysis for that language, phonology articles average a per-language depth of 19% (meaning: for each language used in a phonology article, on average that language constituted 19% of the total linguistic data in the article), but syntax articles run twice that at 39%.

In syntax articles, 63% of the languages cited made up at least 2/3 of the data in the article which they appeared in, so that in the article we could say that the language is the focus language of the article—this corresponds to the intuitive observation that syntax articles generally give a detailed analysis of a phenomenon in a language. By contrast, in phonology articles only 11% of the languages

are discussed at comparable depth. In the phonology articles, the 17 languages which achieved this 66% or higher depth-index were the featured language in exactly one article. 35 languages get an equivalent focus in syntax, and frequently one finds the same focus-language being discussed in multiple articles, thus we find Chinese being the focus of 8 articles (i.e. constituting at least 2/3 of the data in the article), and English was the focus of 21 articles.

(19) Focal languages

a. Phonology

Bengali, Columbian Salish, Cupeno, English, Indonesian, Kashaya, Latin, Macedonian, Maltese, Mohawk, Mongolian, Nawuri, Polish, Shanghainese, Shona, Slovak, Yawelmani (1 article)

b. Syntax

- Balinese, Bambara, Chamorro, Chichewa, Greek, Hungarian, Korean, Marathi, Mohawk, Moroccan Arabic, Palauan, Persian, Portuguese, Selayarese, Turkish, Tzotzil, Welsh (1 article)
- Catalan, German, Haitian-Creole, Hindi, Japanese, Polish, Spanish, Yiddish (2 articles)
- Icelandic, Russian, Italian, Breton, Dutch, Irish, French, Modern Hebrew (3-7 articles)
- Chinese (8 articles), English (21 articles)

Thus the depth of empirical coverage in syntax is much greater, even though—or perhaps better said exactly because—there is comparatively less variation in the range of languages that are looked at.

A last set of revealing statistics pertains to the source of the data used in articles. Another property of articles noted in this survey was whether the data come directly from a native speaker, either one of the authors or an informant, or from a secondary source such as a grammar book or another article written by a native speaker or

fieldworker who personally gathered the data. The first set of figures indicates the percentage of languages in articles whose data come from secondary sources. The second set of figures indicates the percentage of data which comes from secondary or lower sources (e.g., citing data from an article where that author did not gather or generate the data). The difference in these two figures for syntax vs. phonology reflects the fact that syntax articles tend to make relatively minimal one-example use of secondary data and thus secondary data is more marginal to syntactic articles, whereas secondary data is more central to phonological research.

(20)

	Phonology	Syntax
Languages with secondary data sources:	88%	48%
Percentage secondary data:	61%	11%
Percentage of native speakers:	14%	40%

In only 4 phonology articles were the authors (or one of the authors) native speakers of the languages discussed, namely Japanese, English, Polish, and Bengali. In syntax, only 48% of the languages are cited from secondary or lower sources, and those examples account for only 11% of the syntax data. 40% of the languages investigated are the native language of one of the authors.

What could explain this fundamental difference between syntactic and phonological research, in terms of crosslinguistic depth and breadth of hypothesis testing? Existing documentation in published general-purpose grammars is rarely very good for syntax, especially for research questions that require sustained and in-depth knowledge of sentence structure which carefully controls variables. Most work in syntax by its nature demands access to a native speaker, especially when crucial data involves the difference between the analogs of “Who does Bob believe Tom to have explained the solution to?” and “What does Bob believe Tom to have explained to Fred?”. For this reason, syntactic researchers are most likely to work on their

own language because they have instant access to the relevant data; they are second most likely to work on a language spoken by a linguistic colleague or student who speaks the language (thus a potential co-author) for the same reason; and least likely to work on a language using a regular language informant, because it can take very many years to get a solid grip on the myriad complexities in the syntax of a language, and often a naive informant lacks the crystal clear syntactic data judgements that a trained linguist has.

Book data is more likely to be at least reasonably suitable for the questions asked in phonology, even if it is less than ideal. For a phonology article, the data which can typically be extracted from a published grammar is often spotty, and very commonly, phonologists trying to present the details of a language's structure are thwarted by the poverty of available data since gramar-writers do not always provide extensive paradigmatic data on word formation especially as it involves phonological alternations. Usually you cannot make an absolute knock-down argument for a particular analysis of a language with extensive documentation, and you can't discuss all of the complications and contingencies: but you can make a reasonable case for a particular claim. That means that it is easier to come up with the minimum required amount of data for a phonology article using only previously published sources than it is in a syntax article. This is the main reason why phonology articles tend to have more superficial coverage of individual languages—published sources provide less information that native speakers. This also explains why they can cover a wider range of languages: books are a much more convenient source of data than native speakers are.

4. Conclusions

The central problem in establishing universals of language is establishing a proper empirical basis for claiming that the property is a property of the abstract entity 'human language' in the general

property of the abstract entity 'human language' in the general sense, and is not just accidentally true of a select assortment of specific languages. A large and diverse crosslinguistic database is therefore a sine qua non for empirically justifying claims of universality. Since most languages in the world are almost completely underdescribed, and since under-description affects vast geographic spans and linguistic genetic groups, it is very hard to argue that universal linguistic claims have been truly adequately tested. The main desideratum for development and testing of empirical universal hypotheses is thus a concerted program of in-depth language description.

Technological and socio-economic changes are having a huge impact on cultural diversity throughout the world. The existence of hundreds of very small languages in New Guinea or parts of Africa is basically a consequence of historical isolation, where people living in a remote area have relatively little contact with other people fifty miles away. Villages in the contemporary world are becoming less isolated thanks to motorized transportation and road-building, and more and more people are moving into major urban centers dominated by a single language. Even in small villages, frequent indirect contact with the major languages via radio and television is destroying the linguistic isolation which has over the millennia allowed the independent development of thousands of different languages. The cultural and linguistic diversity of the world is rapidly diminishing, to the point that Michael Krauss estimates that 90% of current languages will be dead in a century. This prognosis may seem bleak, but it is important to remember that most of the linguistic diversity of the world is in the hands of very small communities. It is unlikely that the 250 Oneida, 40 Yukaghir, or last 6 speakers of Quinault will pass their languages on for many more generations.

With far greater than chance probability, a language currently spoken with fewer than a million speakers will be dead within a century or two. With far greater than chance probability, languages which are currently undocumented will remain undocumented in the

future. With far greater than chance probability, the majority of language will die out essentially undocumented. With far greater than chance probability, a given theory of linguistic universals which is actually wrong will have the appearance of correctness because the counterexamplifying languages died out undocumented.

References

Archangeli, D. 1985. Yokuts Harmony: Evidence for Coplanar Representation in Nonlinear Phonology. *Linguistic Inquiry* 16, 335-372.

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

Chomsky, N. & M. Halle. 1968. *The Sound Pttern of English*. New York: Harper and Row.

Clements, G. 1985. The Geometry of Phonological Features. *Phonology* 2, 225-252.

Clements, G. & E. Hume. 1995. The Internal Organization of Speech Sounds. In J. Goldsmith (ed.), *The Handbook of Phonological Theory* 245-306. Oxford: Blackwell.

Comrie, B. 1981. *Language Universals and Linguistic Typology: Syntax and Morphology*. Oxford: Blackwells.

Dimmendaal, G. 1998. A Syntactic Typology of Surmic from an Areal and Historical-comparative Point of View. In G. Dimmendaal & M. Last (eds.), *Surmic Languages and Cultures* 35-81. Nilo-Saharan Linguistic Analyses and Documentation vol 13. Köln: Rüdiger Köppe.

Dryer, M. The Greenbergian Word Order Correlations. *Language* 68, 81-138.

Greeenberg, J., C. Osgood & L. Jenkins. 1963. Foreword to J. Greenberg. In J. Greenberg (ed.), *Universals of Language* 25-51. Cambridge, MA: MIT Press.

Greenberg, J. 1963. Some Universals of Grammar with Particular References to the Order of Meaningful Elements. In J. Greenberg (ed.), *Universals of Language* 73-113. Cambridge, MA: MIT Press.

Payne, T. 1997. *Describing Morphosyntax*. Cambridge: Cambridge University Press.

Ross, J. R. 1967. *Constraints on Variables in Syntax*. Doctoral dissertation,

Cambridge, MA: MIT.
Sagey, E. 1986. *The Representation of Features and Relations in Nonlinear Phonology*. Doctoral dissertation, Cambridge, MA: MIT Press.