Mathematical Methods in Historical Linguistics: their use, misuse and abuse

by
H.M. Hubey
Associate Professor
Department of Computer Science
Montclair State University
Upper Montclair, NJ 07043

hubeyh@montclair.edu
http://www.csam.montclair.edu/~hubey

Contents

0. Reasoning Methods
1. '*Randomness*' and Density Functions
2. *'Regularity'* of Sound Change and the Birthday Problem: the Pigeonhole Principle
3. Binomial and Poisson Densities
4. Bayesian Reasoning; why can't we use it?
5. The Comparative Method and Diffusion
6. Lexeme Loss, Glottochronology and Stochastic Differential Equations
7. Conclusion

Appendix A: Solution of the False Cognacy Problem
Appendix B: Examples of Uses of Bayesian Reasoning
Appendix C: Some Probability Mass Functions and Densities
Appendix D: Distance and Phonetic Similarity
Appendix E: Measurement Scales
Appendix F: Differential Equations and Carbon Dating
Appendix G: Probability Theory Foundations
Appendix H: Phonological Phase Space

References

Mathematical Methods in Historical Linguistics: their use, misuse and abuse

by
H.M. Hubey

Abstract

The language change, and diffusion of semantic and phonological characteristics of languages is modeled mathematically. It is shown that the various probability densities can easily masquerade as others and that fitting some data points regarding sound shapes of words of languages to a particular probability density does not justify the conclusion that the languages are not genetically related. It is shown that lexeme loss itself is a random process which is comprised of both a deterministic and a random component.

The correct methods to use for reasoning about language evolution, change and geneticity is shown. Relationships of these to glottochronology and other simple methods is shown. Examples are given of Bayesian reasoning and the use of Fokker-Planck-Kolmogorov methods to obtain rigorous results in historical linguistics. Problems of measurment, distance, similarity, and specific methods of measurement of phonological distances is shown. Measurement scales are an extremely important topic in all social sciences since they deal with attempts to produce metric spaces (mathematical spaces with distance defined over them), and should also figure prominently among the basic issues of historical linguistics. The production of family trees (classification) should rightly follow from and after the solutions of the issues of measurement.

A complete and rigorous solution of the “false cognacy” problem is given using the inclusion-exclusion principle and also recursive functions, and examples are given. In short, it shows what kind of reasoning not to employ and which kind to employ along with lots of examples, gives a rigorous solution to the main problem of historical linguistics at the present level of rigour and knowledge, and also shows how to obtain more exacting and more rigorous solutions, and is thus a program of research for the future. It is therefore at once, an almost complete tutorial on mathematical and scientific methods to be used in historical lingustics, a research paper with solutions to the most basic problems of historical linguistics and a research programme for the future whose absence is not even noticed presently. A Gedanken experiment for the Oswalt method shows why and how it might work. The many appendices and examples all deal with peripheral phenemena with which historical linguists should be very familiar.

0. Reasoning Methods

> "It is not always clear whether the similarities observed between the lexica of different languages could easily be a result of random chance or must reflect some historical relationship. Particularly difficult are cases in which the relationship posited is remote at best; such cases must be evaluated by comparison with mathematically valid models whcih realistically simulate chance resemblances between languages. A simple but relatively detailed method for making such comparisons has been described and exemplified in Ringe (1992, 1993) but more general comparisons are also useful."

These lines belong to Ringe [1995] in his welcome attempts to inject rigor into historical linguistics. However there are very serious errors in the way in which Ringe's "methods" are being used. There is much confusion in the literature on the usage of probabilistic methods in diachronic linguistics. This paper an attempt to point to the correct use quantitative methods in historical linguistics, as the others Hubey[1994] [Hubey,1998a,b], Hubey [1999a] and to correct the misuse of probabilistic methods to reach conclusions not backed up by probability theory.

There seem to be several types of attempts to use probability theory in the literature:
1. statistical
2. quantity due to chance
3. chance distribution

By statistical I mean work done by Nichols, works that can be seen in Embleton, or correlation-regression analysis as done by Labov, and this paper has nothing to say about any of these. The second category of works attempt to determine "how many" putative cognages (PCs) we can obtain due to chance alone. Among these are the types that attempt to use the binomial density as in Ringe[1992], and those that calculate the probability of CVC syllables and the like to seem to be cognates due to chance [Bender,1969], [Cowan,1962], [Swadesh,1954], Greenberg [1960]. In some unpublished works presented on web pages, we find calculations (based on the binomial density) that up to 400 false cognates may be obtained due to chance. The third type are those that attempt to argue that if the distribution of cognates resembles that of the binomial density then it must be due to chance [Ringe,1995]. A fourth way to attempt to obtain "ball-park" figures on the number of putative cognates (PCs) due to chance is in Hubey[1994].

1. Randomness, Density Functions and Ringe's Work

It is possible for a process said to be random to consist also of deterministic components. This seems to be one of the sticking points of the arguments being waged for and against Nostratic. Furthermore, the process of loss of lexemes is random. Therefore if the original lexicon is given by the set O and the lost lexemes are represented by the set L, then what is left which is the set difference O-L itself is random because the set L is random. An example of incorrect usage is where he makes statements such as "*I will show that the distribution of Nostratic cognates over the six proposed subgroups of the family appears to be a random distribution of the type just described*" Ringe[1995:56]. In other words, if the loss of lexemes is due to random processes, and is therefore random, then the lexemes that are left in the language is also random. Furthermore, the comparison to the binomial distribution seems to be unmotivated.

In his attempt to show that Nostratic cognates are due to chance, Ringe [1995] shows that the distribution resembles that of the Binomial. Apparently the train of thought here is that if any set of data can be fit into some probability density, then it must be due to chance and that no deterministic relationship of any kind exists. Perhaps this is too strong a statement but it is not exactly clear what is meant by fitting the data onto a probability density. However there are many places in which he implies that this is indeed his reasoning. For example, he writes;

> " *These results are astonishing in at least two ways. In the first place, it is remarkable that the distribution of putative cognates approximates so closely curves generated by random distributions of the most straightforward kind; even if all the cognates adduced are illusary, and their similarities are the product of sheer chance, one would expect the cluster of chance events that gave rise to those similarities to be more complex. Secondly, it is even more remarkable that the probabilities whose curves the distribution of putative cognates approximates are so high (.4 and .7).*"

First, there is nothing to be astonished about. Secondly there is nothing straightforward about the Binomial density because there is nothing crooked about any other probability density. Thirdly, there is no reason to assume that the clusters or the chance events that gave rise to these to be "more complex". More complex than what? On what basis is this to be expected? and fourth, why is it surprising that the probabilities are "so high"? How high should they have been?

There are many reasons why conclusions of the type hinted at by Ringe cannot be entertained;

1. deterministic chaos
2. determinism + randomness = randomness
3. determinism*randomness = randomness

First of all, it is now known that chaos (which is random-like) can be produced by deterministic means. Secondly, if randomness is added to a deterministic process the result is a stochastic (random) process. So finding randomness in data does not imply that there is no determinism. On the contrary, we expect everything we measure in this world (i.e. the data we collect) to be contaminated by "noise" (i.e. error), and it is these errors that are said to be "random". Indeed computational deterministic chaos is about the spread of error throughout the iterative computation. Noise or randomness can also be injected into a deterministic system via multiplication. Finally, the loss of lexemes and sound changes are said to be random. If that is the case, then the set of all lexemes lost is a random set. Subtracting this set from the original lexicon then results in another random set, the set that was not lost and is presumably what Nostraticists are finding albeit with more corruption with noise. So we expect the data to fit into some kind of a pattern. And this is a very common problem with students learning probability theory. They expect randomness to mean that there is no pattern. That is not true at all. There are lots of patterns in random data. Probably the easiest and most convincing demonstration of this fact is the Sierpinski Gasket, which is produced by the so-called "chaos game" as dubbed by Barnsley[1988] The rules of the game are simple. Put three points on a piece of paper so that they roughly approximate the vertices of a triangle, and put a point at random on the paper preferably inside the triangle (but it does not matter). Then throw a die. Depending on the values (say, {1,2} for point A, {3,4} for point B, and {5,6} for point C) measure the distance from this original point to the appropriate vertex and put a dot halfway between this vertex and the original point. Now repeat this process of tossing a die and putting a dot halfway between the vertex and the last point drawn. One would expect a huge mess of points

cluttering the paper at random. On the contrary one gets a beautiful and quite ordered set of points. As another example of the fact that there is no reason the seemingly random data precludes geneticity consider this; Let the sequence $O = \{50, 51, 52, 53, \ldots 100\}$ be a deterministic set of integers from 50 to 100. Let $L = \{1, 10, -3, -10, 9, 6, 4, \ldots\}$ be a set of random integers in the range from -10 to +10. Then the sequence/set $P = O - L = \{49, 41, 55, 63, 63, 48, \ldots\}$ is a random sequence of integers albeit with a strong deterministic trend. Therefore if O=the set of

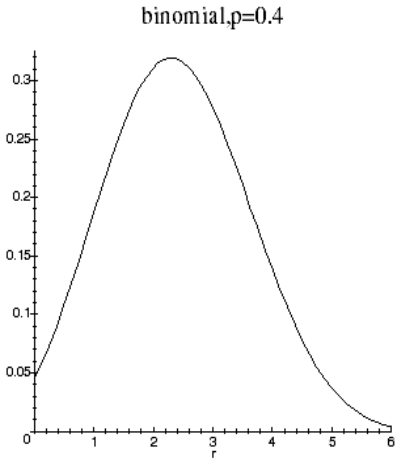


Figure 1a: The Binomial density for p=0.4 as used in Ringe[1995].

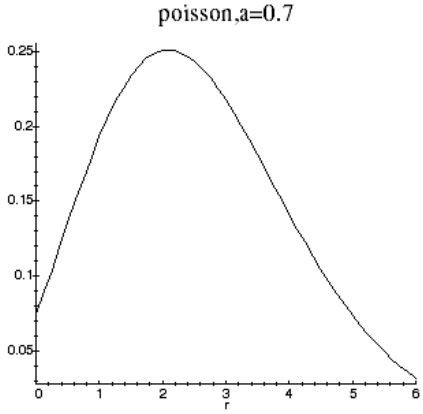


Figure 1b: The Poisson density with the parameter *a* chosen to make the density look like the binomial density in Figure 1.a. The parameter *a* was selected via eyeballing.

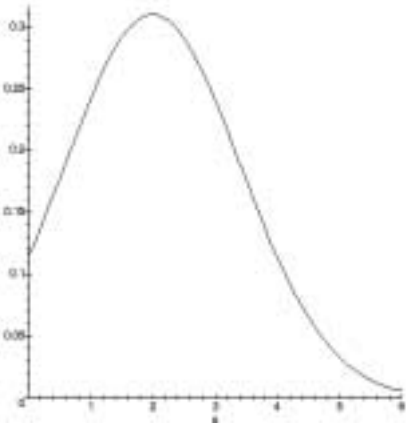


Figure 1c: Gaussian Density with $\mu \approx 0.2$ and $variance \approx 2$.

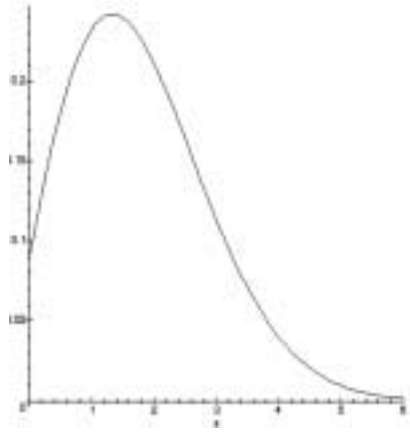


Figure 1.d: Shifted Weibull Density

$$p(x) \sim (x + 0.4)\exp\{(x + 0.(4))^2\}$$

original lexemes of a language, and L=the set of lexemes lost at random, then the set P is the set of lexemes still left in the language, and it is a random set, so this idea cannot be used to prove that the languages in question are not related.

If the word losses in language occur at random, where some of them become replaced by words copied from other languages, and others are derived from other roots/stems, then it should not be surprising that the set of words that are left over at present time do resemble random sets. Indeed, the fact that the data seem random is to be expected. What kind of distribution the data fit can itself be important for other reasons, but the fact that the data itself fits into a probability density certainly cannot be significant in the way in which it is implied in Ringe's works. The data can probably just as easily fit into the Poisson distribution or a Weibull-like distribution. These are plotted below for various values of the parameters and compared against the binomial for the value of p=0.4 which is used by Ringe. Why and how these can or cannot be used is discussed in the appendix on probability densities.

There are more serious problems in the application of the method as well even if none of the above were true (and they are not false). For example, Ringe assumes that we have to subtract out some baseline effects if we find that the number of sound occurences in the first place is not that of being beyond what already exists in the language [Ringe,1992]. If this is the case then if we compared two sets of the basic vocabulary of the same language to itself we would have to subtract out the various proportions of the relevant phonemes ostensibly to account for randomness or due to chance, so the method cannot be correct. In other words, two sets of words from two languages cannot resemble each other unless they have similar sets of phoneme structures anyway, but then Ringe wants to punish the languages for having similar sets of phonemes by subtracting out this as a random/baseline correlation. In this case, comparing the same language to itself using Ringe's reasoning would lead us to conclude that the language is not related to itself. This is really something that is done in Ringe[1992]. Secondly, there is no reason to fit the data to the binomial density since it is for independent trials, but if we do create set of putative cognates, we can use each word only once so that it cannot be used again. Therefore this is an approximation at best. The data could have been fitted a Poisson density or even a Weibull-like density as well.

Thirdly, the binomial density is rarely used in calculations because of the difficulty of calculating factorials. Instead two other densities, the Gaussian and Poisson are used as approximations. So using this reasoning if the data resembled the Gaussian or the Poisson or any probability density we'd have to conclude that it is due to chance. But lexeme loss and sound change are also random processes according to the prevailing beliefs and theories of linguistics. So then fitting probability density to data is simply saying something about the specific behavior of the lexeme loss and sound change process and nothing more. More on this will be shown in the sections below.

Ringe writes yet in another place [Ringe,1995:60];

> "The individual events contributing to to such an acceptance were undoubtedly more complex, and the probabilities that particular events would occur might have varied considerably; but if the overall probabilitiy of an acceptance for a root of more than two consonants was 0.4, the result would indeed be the curve that we find."

Here Ringe seems to lay the ground work for the kind of an argument that is called the *affirmation of the consequent fallacy* [Hubey,199b]. In other words if $P \Rightarrow Q$ (where the sign $\Rightarrow$ is the logical 'implication') then it must also be true that $Q \Rightarrow P$, although he does not say so explicitly.

But if $P \Rightarrow Q$ is true that is equivalent to $\bar{Q} \Rightarrow \tilde{P}$ not $Q \Rightarrow P$. In other words, Ringe seems to be saying "if X happened, then we-find-this-curve" therefore, "we-find-this-curve, therefore X must have happened". Even if he is not arguing this way, it is written in a way in which at least some readers can understand it in this manner. His reasoning also seems to remind one of a kind of reasoning, if it can be called that, which Peirce called "abduction" as can be seen in Lass (1998: 334)

> Rule: all beans from this bag are white.
> Result: these beans are white.
> Case: these beans are from this bag.

Using logic this conclusion is false. However, there is a probabilistic reasoning (statistical inference) which superficially resembles this. It goes something like this:

> Fact: the mean weight of beans in bag A is 40 gm and that in bag B is 60 gm.
> Fact: this bean weighs Y gm.
> Conclusion: with z% certainty this bean is from bag X.

In order to reach such conclusions the distributions (density) of bean weights in both bags must be known. Secondly, it must be known in advance that the bean in question is definitely from one of the bags. Even with these caveats this only superficially resembles abduction. In Ringe's case, his reasoning is definitely not logical, and it does not fit into the statistical inference scenario. So then what kind of reasoning is it? All is not lost. There is still yet apparently a third way of doing science, as done in physics, and its derivatives such as chemistry and engineering which Ringe's reasoning resembles, and that is one of model-building. Of course, by this I mean "mathematical model building" because there is really no substitute for it. However, we should also not forget:

> The advantages of mathematical or analytical models ...are well known.. This does not mean that models formulated in ordinary language (e.g. verbal descriptions) are to be despised. A verbal model is better than no model at all, or a model which because it can be formulated mathematically falsifies reality. Indeed, theories of enormous influence such as Darwin's Theory of Selection, were originally verbal. Much of psychology, sociology and even economics today is still descriptive. Models in ordinary langauge, therefore, have their place in system theory. The system idea retains its value even where it cannot be formulated mathematically, or remains a "guide" rather than being a mathematical construct. Hubey[1979:5]

In another place Ringe attempts to 'explain' the 'discrepancy';

> "Illic-Svityc's sound correspondences are not always sufficiently precise; this is clear from the large number of alternative reflexes listed in his tables ... Moreover his judgement of irregularities in putative cognates is relatively 'forgiving'; he does not, in fact, insist on absolute regularity of sound change. This substantially increases the probability that chance resemblances will be accepted as significant (Ringe:1992:67-70). Irregularities of the second sort, at least, can be tolerated (in suitably small numbers) in working with languages whose relationship has already been established beyond doubt, but they are a potentially fatal weakness in an attempt to prove a doubtful relationship. The same criticisms can be leveled against Illich-Svityc's rather lax standards for the semantic fit between comparanda (cf. Serebrennikov 1986, Vine 1991). This is another source of the increased incidence of chance resemblances that might be deemed acceptable (Ringe 1992:64-67), and it, too, can undermine an attempt to prove a doubtful relationship."

There are a number of very serious and solemn pronouncements here and conclusions which cannot be backed up with any kind of mathematical rigor. In other words, it is completely circular because it assumes that a very fatally flawed work (i.e. Ringe, 1992) taken to be a standard which the criticized paper does not meet. There are so many things wrong with the lines above and Ringe's (1992) work that it will take many pages to correct. We should first make a simple attempt at computing some 'ballpark' estimates. We have n "meanings" distributed to n words in every language at random. We select two languages at random. Assuming that the sound shapes of words can be correlated with each other what is the probability that m words will have the same meaning (and the same sound shape)? This is the same problem as that of calculating the probability of k matches between two languages which have exactly the same N phonological words except that the N meanings have been randomly assigned to the N phonological words?

2. Regularity of Sound Change and the Birthday Problem

The Birthday Problem Statement: We have a group of N people in a group, say a class of students. How many students do you need in class to have a 50% chance of at least two students having the same birthday? The simplest rigorous answer can be obtained by the use of the pigeonhole principle stated below:

Pigeonhole Principle: This says that if we have n pigeonholes and n+1 pigeons there must be at least one pigeonhole with 2 or more pigeons.
This means that if we have 367 persons in a group, there must be at least 2 persons with the same birthday. After all, even if we can organize the first 366 persons so that each has a different birthday, the 367th person has to have one of the possible 366 birthdays. This principle applies immediately and directly to possible sound changes to give us at least some guide posts as to what to expect at one extreme. Suppose the set of phonemes in a language is $L_1=\{l_1,l_2,...,l_m\}$ so that $|L|=m$. We have an observation of n *sound changes* of a single phoneme in language L_1 to other sounds in L_2. If the number of sound changes, n, is greater than m (the number of phonemes) then there must be at least one sound change that has more than one example. If "regularity" of sound change is defined as "more than one example sound change per phoneme" then the observed sound changes between L_1 and L_2 are regular. To show this suppose there are 20 consonants in the two languages in which we are interested. That means that there are 400 possible sound changes which is 380 not including the no-change change i.e. the sound change to itself. This means that if we find 381 sound changes even in the worst case of 380 sound changes each being unique, the 381st sound change has to duplicate one of the other 380 sound changes and thus becomes "regular". However, it is not necessary to have n+1 sound changes to have duplication (regularity). This probabilistic problem is related to the famous Birthday Problem to which we turn now.

As another attempt at solution, suppose we use a Binomial model with 5 vowels and 14 consonants. If sound can change to sound there are 361 possible sound changes, which luckily for us is close to 366. So we can make a quick (and incorrect) calculation using the same idea. We can see that even at around 22-23 alleged accidental matches we will have 50-50 odds that there will be two correspondences showing the same sound change. But this is exactly what the historical method is based on. If we have 50% chance of duplicate sound changes if we find only 23 sound changes, how then does regularity not come into the calculation if 200-300 mathes can occur

purely due to chance? There will a large number of correspondences if we have allegedly 400 accidentaly cognates, and thus we will have "regular sound correspondence". This solution claiming up to 400 accidental correspondences as likely events cannot be correct.

But the standard way in which a typical linguist argues about the *value of regular sound change* is; "Wow, I have 400 putative cognates (PCs) and all regular sound change too!". However, via the Birthday Problem's solution we know that if we do indeed have 400 putative cognates (even if they are due to chance) we expect lots of regular sound changes because regularity is a side effect of quantity. Therefore regular sound change if the number of PCs is large is not significant since it is really the quantity that is important and the regularity is a side effect. The only time regular sound change is significant is if the number of PCs is small. But if the number of PCs is small, typically it will be ignored, precisely when it is important.

Let's first try a simple method for the birthday problem which is incorrect. The probability that any two days selected from 366 days is the same day is 1/366. Let this be the two-way matching probability, p. Suppose we have 3 dates, x,y,z.? Then we have the possibility of x:y, x:z or y:z matching. Therefore we should multiply p by 3. Similarly if we have n dates at random, then we should multiply p by the number of edges in a compete graph K_n since that is the number of possible multiple matchings. For example, for n=23 we would multiply the two-way matching probability by (23)(22)/2 which is (1/366)(23)(22)/2 = 0.69125. Since the correct answer is approximately 0.5, it is about 40% off.
The correct solution: If we have k possible dates and we have n persons in the group then the probability that they all have different birthdays is

2.1) $$P(n\ \ different\ \ birthdays) = \frac{k}{k}\cdot\frac{k-1}{k}\cdot\frac{k-2}{k}\cdot\ldots\cdot\frac{k-n+1}{k}$$

The first thing to notice is that this is not the binomial density. The binomial density can be used if the trials are with replacement meaning that after we have a match, we put them back and look for another match with the whole set of words of both languages, instead of taking out the matches that we found already out of the running. The birthday problem is of this type. In other words, the first person has a choice of k days in which he can be born and not match anyone's birthday. The second has only k-1 days (excluding the first person's birthday) and so on. We can write the first form using factorials

2.2) $$\frac{k(k-1)(k-1)\ldots(k-n+1)}{k^n} = \frac{k!}{k^n(k-n)!}\ .$$

Using factorials for computations is difficult and therefore we would like to have a simpler approximation. Approximating k! by the Stirling approximation as k^k or e^{klnk} we obtain

2.3) $$\frac{e^{k\log k-k}}{e^{n\log k}e^{(k-n)\log(k-n)-k+n}} =$$

$$= e^{(k-n)\log(k)-k}\ e^{-(k-n)\log(k-n)+k-n}$$

$$= e^{(k-n)\log(k)-(k-n)\log(k-n)-n} = e^{(k-n)[\log(k)-\log(k-n)]-n}$$

finally obtaining

$$P(n\ different\ birthdays) = e^{(k-n)\left[\log\left(\frac{k}{k-n}\right)\right]-n} \qquad 2.4)$$

This formula can be found in Hubey[1998]. The solution to this problem can be obtained to be 23 using the long way without using the Stirling approximation. If we now substitute k=366, and n=23, we obtain P(n different dates)=0.477920. So that it is very close to the answer. We can invert this problem and solve for n, by setting the rhs equal to 0.5 (or to any other probability). The problem of what kind of a distribution we can expect if n persons are selected at random is a different kind of a problem.

3. Binomial and Poisson Densities

It is often confusing for those not familiar with the subtleties of probability theory to be able to keep track of the reasoning involved and knowing which formula to use when. In case of language comparison, it is easier if we think of being given N addresses or containers which stand for meanings of semantemes. These are the words of one of the languages being compared. What we want to do is find words that sound similar and which have the same meaning. It is as if we have N numbered containers each with a single ball inside with the same number. We throw up the balls so that each lands in one container at random. We want to know how many of them will land in their own containers. This is also called the hat-matching problem, and has already been solved. An application of this to historical linguistics can be seen in Hubey[1994]. Another way of looking at the problem is to have N numbered addresses. Then we pick numbers at random and see how many balls fall into which container. Specifically, this should resolve the problem of what kind of a probability density should result if the matches (which are alleged cognates) are at random. This is the specific problem which Ringe attempts to solve [1992,1993]. Aside from reaching the wrong conclusions, he also misses the important fact that regular sound change is a by-product of quantity. In other words, if one can find 400-500 cognates then the sound regularity is bound to occur. On the other hand, regular sound change is much more significant if the number of matches (putative cognates) are small. Specifically we want to know such things as;

1) How many words of A will have no matches in B
2) How many of the words of A will have exactly one word in B match it
3) Exactly two [three, four,...] words can be found to be putative cognates.

Assuming a uniform distribution of N words, the probability of a specific word being chosen as a match is P(A)=1/N=p. If two words are selected at random, the probability that both words will be putative cognates for the same word is P(AA)=(1/N)(1/N)=1/N^2=p^2. In general we'd like to know how many words are putative cognates with a certain word without regard to their order. The outcomes are again of the Bernoulli type, where success is matching the given word and failure is not matching that word. Clearly this is the Binomial Distribution given by

$$f(x) = \frac{r!}{x!(r-x)!}p^x q^{n-x} = \frac{r!}{x!(r-x)!}\left(\frac{1}{N}\right)^x\left(1-\frac{1}{N}\right)^{r-x} \qquad 3.1)$$

The expression is clumsy for calculations. Therefore often an approximation is used. The Poisson approximation to the binomial is obtained rather easily.

If $x \ll r$, then $r(r-1)(r-2)\ldots(r-x+1) \approx r \cdot r \cdot \ldots \cdot r = r^x$. Also $1-p \approx e^{-p}$, and $(1-p)^{r-x} \approx e^{-(r-x)p} = e^{-rp}$. Therefore making the substitutions we can approximate the binomial with

3.2) $$f(x) = \frac{r!}{x!(r-x)!}(1/N)^x(1-1/N)^{r-x}$$

with the final result given as

3.3) $$f(x) = \frac{\left(\frac{r}{N}\right)^x e^{-\left(\frac{r}{N}\right)}}{x!}$$

In this specific case

N=the number of words of language A
r=the number of words of language B to be matched as cognates
x=the number of words of B assigned as a cognate to a given word in A

With N=r the probability that no words of B will match a given word of A is

3.4) $$f(0) = \frac{1^0 e^{-1}}{0!} = 0.368 \; ; \; f(1) = \frac{1^1 e^{-1}}{1!} = 0.368$$

For two, three, or four addresses the calculations are

3.5) $$f(2) = \frac{1^2 e^{-1}}{2!} = 0.168 \; ; \; f(3) = \frac{1^3 e^{-1}}{3!} = 0.061 \; ; \; f(4) = \frac{1^4 e^{-1}}{4!} = 0.015$$

As can be seen there is a very rapid drop-off. In general if there are N words in A then the expected number of words of A with x words assigned to them as putative cognates is Np(x). This means that for a list of 100 words, the number of words of A with x putative cognates is 100p(x). So we will have ~17 words in A with 2 putative cognates each, ~6 words with 3 putative cognates each, 1 word with 4 putative cognates each, etc.

We can twist other problems into this shape so that we can solve them. A linguist claims that there can be up to r accidental cognates between two unrelated languages. Suppose that there are only N possible sound changes possible between these two languages. What is the probability that exactly x sound changes are of the first type of sound change? This is the problem of distributing r items to N boxes and we want to know the probability that x of them will be assigned to a given (same) address. Exactly the same problem occurs when hashing functions are used to distribute r records to N places, and when we want to know the probability that x records will be hashed to the same address. This is the binomial mass function, again

3.6) $$P(r, N, x) = C(r, x)\left(\frac{1}{N}\right)^x\left(1-\frac{1}{N}\right)^{r-x}$$

where $C(r,x) = \frac{r!}{x!(r-x)!}$. Note that for r=300, and N=221 we can compute x=0,1,2,3..

3.7) $$P(300, 221, x) = C(300, x)\left(\frac{1}{221}\right)^x\left(\frac{220}{221}\right)^{300-x}$$

It is much easier if we use an approximation for this case. Since 1/N is small we can use the Poisson distribution.

At the other extreme we can examine what would happen to whole languages, albeit very primitive languages, such as those that might have been spoken by our ancestors 1-2 million years ago. How many ways are there to distribute k semantemes into n phonological words with exclusion (meaning that a phonological word may have more than one meaning). This is a problem of forming a permutation of size k, with unrestricted repetitions, taken from a set of size n. Therefore there are n^k different ways to distribute k distinguishable semantemes into n distinguishable phonological words without exclusion. In other words, if a primitive language during the evolution of humanoids had to have the power to represent (distinguish) about 100 different meanings (semantemes) by 100 different phonological words (i.e. distinguishable sounds), then there would have been 100^{100} ways to do this if there is no correlation between the sound shape of the word and its meaning. This number is $10^{2^{100}} = 10^{200}$. As a way of comparison, the total number of elementary particles in the universe including photons is calculated to be 10^{87}. The result is that if a group of chimpanzees who have learned about 100 sounds (or symbols) and have associated them with concepts happen to have exactly the same sounds and meanings, the chances are almost zero that they developed independently or each other or that it is due to chance.

A related problem is that of sampling from languages. As an example, let us consider a batch of 100 alleged cognates (discovered by an amateur) checked by a professional linguist who examines 10 of them at random. If none of the 10 cognates is a false-cognate, he accepts the whole batch. Otherwise the whole batch is subjected to further inspection. What is the probability that a batch containing 10 false-cognates will be accepted as a batch of 100 true cognates?

The number of ways of selecting 10 items out of a batch of 100 items is C(100,10). By hypothesis these combinations are equally probable; the items are being selected at random. Let E be the event that "the batch of alleged cognates is accepted by the linguist". Then E occurs whenever all 10 items belongs to the set of 90 true cognates. Hence the number of combinations favorable to E is N(E)=C(90,10). Then it follows that the probability of E is

3.8) $$P(E) = \frac{N(E)}{C(100, 10)} = \frac{90!90!}{80!100!} \approx \left(1 - \frac{1}{10}\right)^{10} \approx \frac{1}{e} = 0.3678$$

The next section shows a more refined and methodological approach called Bayesian reasoning as applied to linguistics.

4. Bayesian Reasoning: Why Can't We Use It?

Statements of the form "About 50% of the languages are agglutinating, so we cannot use it for genetic comparisons" or "There is nothing about agglutination that is genetic" are often found in discourse on historical linguistics methods. First, the fact that 80% of the world's humans are white is not a statement about the fact that whiteness is genetic and is inherited from parents. Secondly the fact that some languages seem to change typology is no reason to form conclusions that it is not inherited. Indeed speakers of agglutinating languages like Turkish inherited it from their parents and they will also pass on the agglutinating characteristics of the language to their offspring. What people often mean by statements of the type above can be better appreciated after examining some simple cases which might at first seem to superficially resemble the inheritance and inheritability of various characteristics of languages.

From the addition and multiplication rules/theorems of probability theory (please see Appendix F), it is possible to derive the Bayes Theorem which states that

4.1) $$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{\sum_{k=1}^{k} P(H_i)P(A|H_i)}$$

Bayes formula can be derived as follows. Let the sample space of the experiment be divided into k mutually exclusive regions H_1,H_2,...H_k. These regions represent the k possible causes of an experimental outcome. Let A be the event that occurred when the experiment was performed and consider the problem of calculating the probability that H_i was the cause of occurrence of A. We have

4.2) $$P(H_i|A) = \frac{P(H_i)}{P(A)}$$

4.3) $$P(H_iA) = P(H_i)P(A|H_i)$$

But event A can occur only if one of the possible events H_i occurs. Thus A will occur if one of the mutually exclusive events H_1A, H_2A, ···, H_kA occurs. Therefore

4.4) $$P(A) = P(H_1A) + P(H_2A) + \ldots + P(H_kA)$$

Expanding each term on the right using equation (4.3) gives

4.5) $$P(A) = \sum_{i=1}^{k} P(H_i)P(A|H_i)$$

Substituting this and Eq. (4.3) into Eq. (4.2) gives Bayes Theorem, Eq.(4.1). As an example of the use of Bayes Theorem let us look into a hypethetical situation of a language in which 45% of the words are from Latin. Another 40% are from Germanic and 15% are from Celtic. Out of all these, about 5% of the Celtic words are about body parts, compared to 3% of Germanic and 1% of Latin derived words. If we pick a word which denotes a body part at random, what is the probability that it is from Latin? Let L=Latin, G=Germanic, C=Celtic, and B=body part. Then Bayes theorem says:

4.6) $$P(L|B) = \frac{P(B|L)P(L)}{P(B|L)P(L) + P(B|G)P(G) + P(B|C)P(C)}$$

4.7) $$P(L|B) = \frac{(0.01)(0.45)}{(0.01)(0.45) + (0.03)(0.4) + (0.05)(0.15)} = 0.1875$$

As yet another example of the usage of Bayesian reasoning suppose that we suspect that the speakers of a language L have lived in the proximity of speakers of another language M with probability p sometime in the past. If these two languages were neighbors then the probability that L would have borrowed words from M is 0.8 because it is thought that M was a high-status language. Any word in L is either descended from the language from the beginning, or is borrowed from M. If we find some word in L that seems to correspond to another in M, what are the chances that it was borrowed? We have three choices: descent from the ancestor, borrowing from M or pure accident. Let C=event that we have an apparent cognate. Let B=the event that it was borrowed

4.9) $$(B|C) = \frac{P(C|B)P(B)}{P(C|B)P(B) + P(C|A)P(A) + P(C|D)P(D)}$$

Clearly calculating some of these probabilities is a formidable task. We would need to find some other ways to calculate some reasonable approximations to these probabilities.

5. The Comparative Method and Diffusion

What all of this points to is the main problem involving calculations of this type. We need to have some way of calculating some basic probabilities. While we are doing this we can again think about probabilities of inheriting typology or inheriting morphology or indeed any measurable characteristic of language, obviously, including words (free morphemes). Why couldn't we use Bayesian reasoning and take into account the most basic characteristics of languages, including typology, basic words, and borrowings? Before we look into this in more depth, we should reiterate some of the basic principles of the accepted comparative heuristic. First, we know very well that borrowed (actually copied) words also go through regular sound change or correspondence, so we cannot use it indiscriminately. Therefore there is a secondary principle, really an assumption, that we should use words which have a resistance to borrowing. In order to fully appreciate the real problems of this method we should look at it from a distance with full objectivity. What evidence do we have that this assumption/principle indeed holds true? So far the overwhelming form of the argument seems to be of this form:

1. We know that the languages of IE family are genetically descended.
2. In this family, there are a certain set of words which can be found in almost all of the languages, and therefore they are quite resistant to borrowing/copying.
3. Therefore we can conclude that this set of words is resistant to borrowing and can be used to test for geneticity.

Aside from the obvious, such as the fact that the principle does not seem to hold except for IE languages [Dixon,1977], and examples in which we can find some of these words copied, the argument is quite clearly circular. First of all, in statement 1, we have neglected to state the most obviously glaring circularity; that is, the alleged geneticity of the so-called IE languages is based on the fact that there is regular sound correspondence amongst these languages for many words. But they could have been borrowed. In order to get around this difficulty we have simply created a circular argument that some words do not get borrowed and we have allegedly shown that it is

true by showing that these words exist in the IE languages. So the only real arguments here are no different in quality than the arguments against using other characteristics of languages such as typology. In this we are led to circular reasoning such as “English is isolating and it is Germanic, therefore language typology can change, but the words remain”, but one can easily counter “But these words that remain are really words that are borrowed/copied by speakers of Celtic, and it was the inability of the new language learners that created the isolating English language, so what really happened is that English is really a creole-Celtic language that later borrowed lots of Germanic words.” So the bottom line is that deeper issues have to be settled elsewhere or as a result of more deep thought. This is a form of the Sorites Paradox and is truly difficult. Instead, we have to look into a simplification of the problem. Suppose that the changes to languages such as the theorized propensity of languages to go through the process as shown in Figure 2 did not occur and that typology was also genetically transmitted.

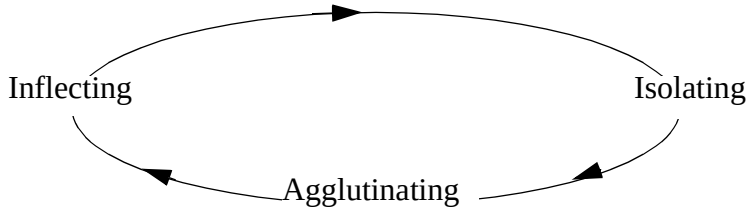


Figure 2 : A Proposed Long-Term Typological Tendency: This tendency would/could interact with the random process of lexeme loss, borrowing, and change to create random process which does not result in uniform distributions.

If this proposed tendency did not exist, and the change of typology was random we could then use typology along with the number of morphemes as cognates as a part of the method to determine language geneticity. Of course, we are treading on really emotional issues here since people normally tend to connect language with race, despite the fact that linguists constantly remind them of the inappropriateness of it. In such a case, then Bayesian logic could indeed be used. However in real life we are forced to deal with the possibility that there are coarse-grained deterministic processes that work alongside others. These are really problems of multiple scales which are of an extremely difficult kind for the present. But the good news is that the existence of such determinism does not leave us powerless. To get an intuitive feeling for the methodology that can be used, we should use the standard “drunk by the pole” analogy. Suppose now that there is a very large platform in the middle of which is a lightpole as before and a drunk executing the same random algorithm. We can still solve this problem. By the word “solve” we mean that we can obtain the probality density so that we can still use these to compute probabilities so that we can use Bayesian logic, meaning that we can take into account typology, and morpheme distributions to reach conclusions about language genes. Problems of this type are solved via methods related to Markov Processes, Fokker-Plank-Kolmogorov methods etc. The simplest such process is the one-dimensional random walk. Before proceeding to other types of stochastic processes, it would be appropriate to derive the Kolmogorov equations for probably the most famous and basic problem in "dynamic" probability theory using very simple methods. Again, this is the "random walk". The usefulness of the random walk problem, aside from developing a "feeling" for probabilistic methods, is that many different stochastic processes can be represented as different aspects of the random walk and it is also a prototypical dynamic stochastic process. As is well-known the con-

cept of prototypes is an important one in cognition/perception and in fuzzy logic. The derivation below is due to Feller [1957:323] and can also be seen in Hubey[1979]. Different and more general derivations can also be found [Stratonovich1967:55-131 or S&V1972:61]. Connections of random walk with other branches of stochastic processes will be shown later. We begin with an unrestricted one-dimensional random walk (a walk in which the probability of a unit forward step is p and a backward step is q) starting at the origin. The nth step takes the particle to the position;

5.1) $$S_n = x_1 + x_2 + \ldots + x_n$$

where S_n is the sum of n independent variables x_i, each assuming the values +1 and -1 with probabilities p and q, respectively.

5.2) $<S_n> = E[S_n] = (p-q)n$ variance of $S_n = <(S_n - <S_n>)^2> = 4pqn$

The time scale can be fixed by denoting by r the (unknown) steps per unit time and by Δ the length of the individual steps. Then ΔS_n, is the distance covered. Thus,

5.3) $$E[\Delta S_n] = <\Delta S_n> \cdot (p-q)\Delta r \text{ and } <(\Delta S_n - <\Delta S_n>)^2> = 4pq\Delta^2 r$$

A probability density function can now be defined such that

5.4) $$P_{k,n} = P(S_n = k) = P(\Delta S_n = k\Delta)$$

Now, $P_{k,n}$ is the probability of finding the particle in the vicinity of k at the nth step and It must satisfy the equation

5.5) $$P_{k,n+1} = pP_{k-1,n} + qP_{k+1,n}$$

The equation above simply states that the particle can come to the kth position on the (n+l)st step either from (k-l)st position going forward with probability p or from (k+l)st position going backward with probability q. Since k is a measure of distance and n is a measure of the time elapsed, equation (5.4) is

5.6) $$\rho\left(x, t + \frac{1}{r}\right) = p\rho(x - \Delta, t) + q\rho(x + \Delta, t)$$

The terms of the equation above can be expanded according to Taylor's Theorem, Using the first-order approximation on the left and second-order approximation on the right (after cancelling leading terms), leads to

5.7) $$\frac{\partial}{\partial t}\rho(x, t) = (q - p)r\Delta \cdot \frac{\partial}{\partial x}\rho(x, t) + \frac{1}{2}\Delta^2 r \cdot \frac{\partial^2}{\partial x^2}\rho(x, t)$$

In the limit as $\Delta \to 0$, $r \to \infty$ and $p \to \frac{1}{2}$ (i.e. the process becomes continuous) in such a way that $(p - q)\Delta r \to k$ and $4pqr\Delta^2 \to D$, the partial differential equation above becomes

5.8) $$\frac{\partial}{\partial t}\rho(x,t) = k\frac{\partial}{\partial x}\rho(x,t) + \frac{1}{2}D\frac{\partial^2}{\partial x^2}\rho(x,t)$$

where the D is the "diffusion coefficient". It is because of this equation, that the forward Kolmogorov equations are known as the Fokker-Planck Equations. The steady-state solution of the differential equation above, is of the form, [Srinivasan & Vasudevan, 1972:55 or Feller1957:326]

5.9) $$\rho(x,t) = \frac{1}{\sqrt{2\pi Dt}} e^{-\frac{1}{2Dt}(x-kt)^2}$$

It should be noted that the value of k is not zero only if the forward-step probability p and the backward-step probability q are not equal. If p=q, then k=0 and the equation above has zero mean so that this models the one-dimensional version of the so-called *drunk-by-the-light-pole* problem in which the drunkard can take a step in any direction with equal probability. In the solution above there is a "drift" meaning that the probability of going in one direction is greater than the other. This drift can come about by other means and can be derived via differential equations which are driven by white noise which is equivalent to the simpler formulation given above.

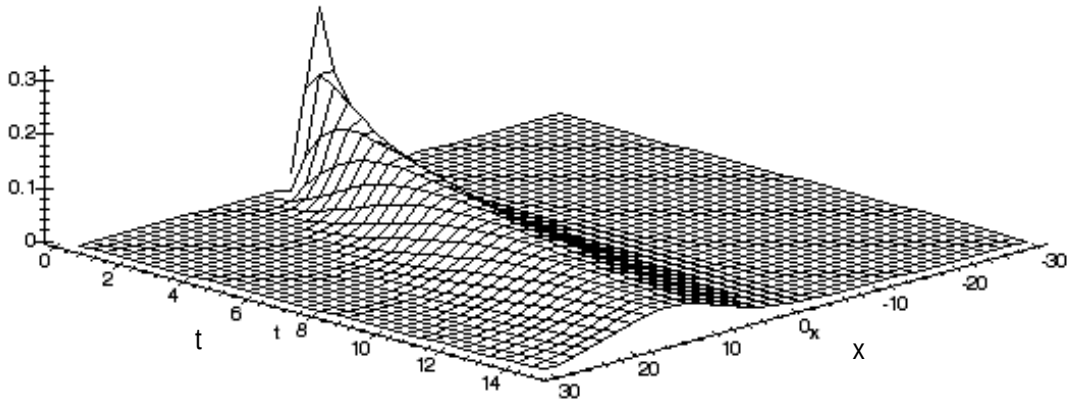


Figure 3: Random Walk (with Drift) Probability Density. The initial probability density (the Initial Condition of the corresponding Fokker-Planck equation) is the Dirac delta function δ(t)δ(x). As can be seen the mean increases with time (moves towards the front in the plot above) so that at t=15 mean is about 15 instead of 0.

It can be seen that the density is a Gaussian. Furthermore, it starts with a highly-peaked distribution (in fact the initial condition is a Dirac delta function) mean-zero density but the mean increases with time because of drift. In addition the variance increases with time so that as $t \to \infty$ the "drunk" (or particle) can be found anywhere. The "drift" came from the determinism inherent in the walk and the white noise (randomness) makes the process random. For the standard "drunk by the pole" problem there would be no drift. That would be equivalent in two dimensions to a heat source placed at the origin spreading out over time and eventually becoming uniformly distributed over the plane. This last distribution is one way in which we can imagine the changes taking place; that is, if we measured the semantic distances d_s, and phonological distances d_p, between the corresponding words in daughter languages over time we would expect it to obey

some kind of a two-dimensional random process since both the meanings and the phonological shapes would "drift" over time away from each other. Now, the jointly bivariate Gaussian density is given by

5.10) $$\rho(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\left(\frac{\frac{x}{\sigma_x}-2\rho\frac{xy}{\sigma_x\sigma_y}+\frac{y}{\sigma_y}}{2(1-\rho^2)}\right)\right\}$$

One often sees the Gaussian bivariate distribution without the correlation in textbooks. As can be seen in the graphs (Figures 4a, 4b), the correlation between the two random variables "stretches" the density along the diagonal so that there will be a greater tendency for the random variables to occur along the diagonal. This fact is very important in what follows.

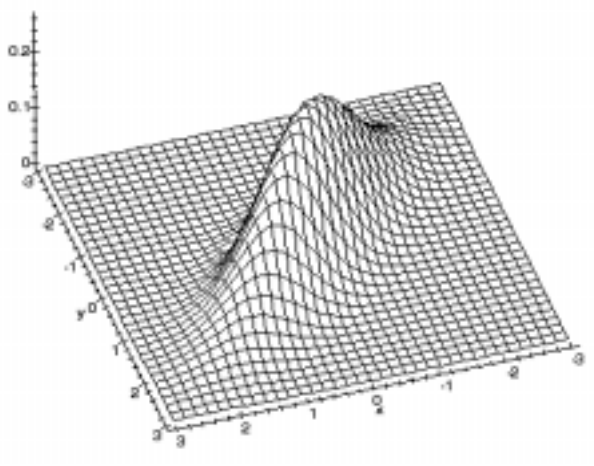


Figure 4a: Jointly Gaussian Bivariate Density with the corelation coefficient ρ=0.8, and σ_x=σ_y=1.

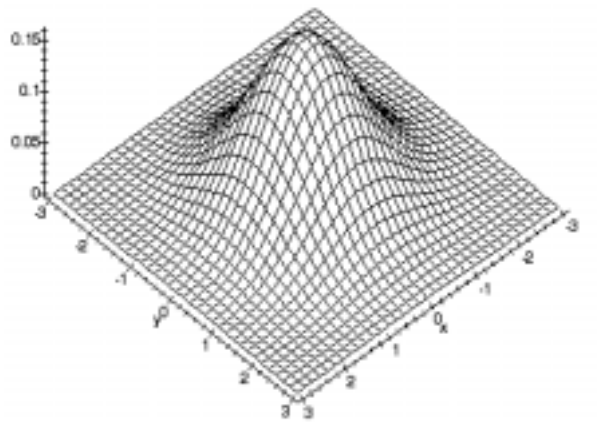


Figure 4b: Jointly Gaussian Bivariate Density with the correlation coefficient ρ=0, and σ_x=σ_y=1.

To treat the case of a random forcing function in a more general setting or put another way, to derive the equation for the probability density for Brownian-motion-like problems (i.e. generalizations of the simple random walk), we need to consider the sample functions or sample solution properties of the stochastic process described as

5.11) $$\frac{dx}{dt} = f(x(t),t) + G(x(t),t)w(t)$$

where w(t) is a Gaussian white noise process which has the formal representation as the derivative wrt time of β(t) which is a Wiener Process where

5.12a) $$\langle w(t)\rangle = 0$$

5.12b) $$\langle w(t)w(\tau)\rangle = \delta(t-\tau)$$

where δ(t) is the Dirac delta function and the triangular brackets <.> indicate averaging. A further specialization of eq.(5.11) is given by

5.13) $$\frac{dx}{dt} = f(t)x(t) + g(t)w(t)$$

This is the starting point of many investigations of Brownian motion. The equation above was first treated by Langevin in his description of Brownian motion for f(t)=constant. Albert Einstein

[1956] obtained the probability density for such a process and thus obtained the first diffusion type equation for the probability density of such a dynamic process. Fokker and Planck later generalized these results. Still later Kolmogorov gave a more rigorous mathematical exposition of the process and also derived the backward Kolmogorov equation. In more general form the Fokker-Planck Equation or the forward Kolmogorov equation for the general linear stochastic process (5.11) is given by (Srinivasan &Vasudevan [1972:55] or Stratonovich[1967:62], Hubey[1983])

$$\frac{\partial}{\partial t}\rho(x,t) = -\frac{\partial}{\partial x}([f(x,t)\rho(x,t)]) + \frac{1}{2}\frac{\partial^2}{\partial x^2}[g^2(x,t)\rho(x,t)] \tag{5.14}$$

There are two other well-known diffusion processes which are particular cases of the Fokker-Planck Equation. If $g^2(x,t)$=constant=2D and f(x,t)=0, the result is

$$\frac{\partial}{\partial t}\rho(x,t) = D\frac{\partial^2}{\partial x^2}\rho(x,t) \tag{5.15}$$

The above equation is a Wiener-Levy process. In addition to Brownian motion, it also represents heat conduction. It was also derived by A. Einstein [1956:15] through physical considerations. The solution is given by eq. (5.9) with k=0. If the coefficients are such that the Fokker-Planck Equation is transformed into

$$\frac{\partial}{\partial t}\rho(x,t) = -\beta\frac{\partial}{\partial x}(x\rho(x,t)) + \frac{\partial^2}{\partial x^2}[D\rho(x,t)] \tag{5.16}$$

then we have, what is called an Uhlenbeck-Ornstein process [Soong,1973] or [Jazwinski,1970] or [Gardiner,1985]. The method of representing a random process by finding the differential equation for the transition intensity ρ(x,t) which must be of a diffusive type is called the Fokker-Planck method and it has been used successfully in many problems in physics, engineering [Hubey,1983], finance [Hubey,1999b], Benninga[1998], Bryis, et al [1998], and biology, Berg [1993]. In order to use these concepts we need to employ the concept of distance (see appendix), and create "instruments" with which various characteristics of languages can be measured. One of the simplest is the "typology" of language. Instead of using only discrete classification schemes, we might want to use the interval [0,1] or [-1,+1] to meausure what might be called the "fusion index". If we produce more such indices so that every aspect of language can be measured then we can hypothesize various types of deterministic changes using equations then use the Fokker-Planck methodology or the related discrete Markov process theory to compute probability densities for the process of language evolution and change. Once such probability densities are available then we could use them in Bayesian inferencing schemes. However, extrapolating from the knowledge of the diffusion equation, the way we observe phonological and semantic distances between cognate words increase over time, and the fact that the words seem to be lost at random we can imagine what we would obtain if we were able to trace a set of words in a single language over time after the language split into a pair of daughter languages.

As can be seen, in Figure 5, the pattern would be that which can be obtained from a probability density which is appropriate for the process and which can be derived from the deterministic laws of change with some noise added using the Fokker-Planck-Kolmogorov methods. As time increases both the semantic distances and phonetic distances of the original words in a single language drift apart in the two daughter families.

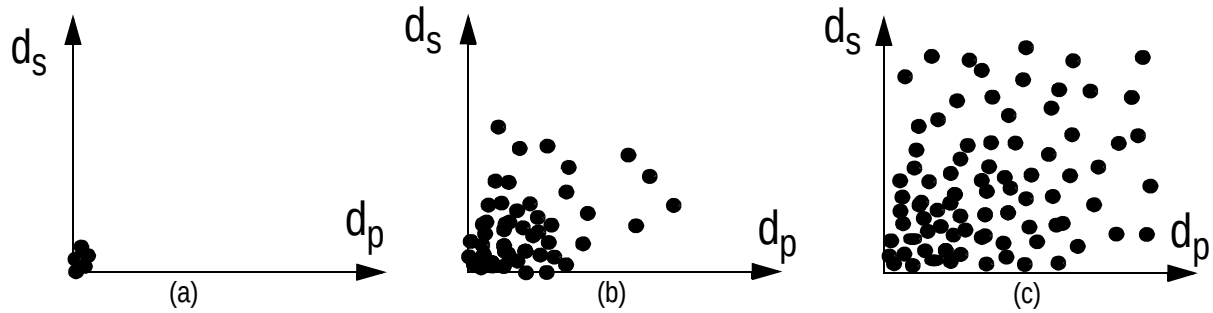


Figure 5: Diffusion of Semantic Distance (d_s) and Phonetic/Phonological-Distance (d_p) over Time: (a) at time t_0 (b) at time $t_1 > t_0$; (c) at time $t_2 > t_1$

This is a graphic and discrete description of the diffusion process. The probability densities connected with this process are shown separately. If a highly concentrated heat source placed at the origin were removed at the initial point in time, it would spread out over time to distribute itself uniformly over the plane. The pictures above are analogous spreading of words from a single language into daughter languages such that over time the distances between both the meanings and the phonological forms gets larger and larger. As can be seen, the fact that some words seem to fit a particular pattern cannot be used to argue that what we observe is mere coincidence and that there is no deterministic component or that there is no geneticity involved.

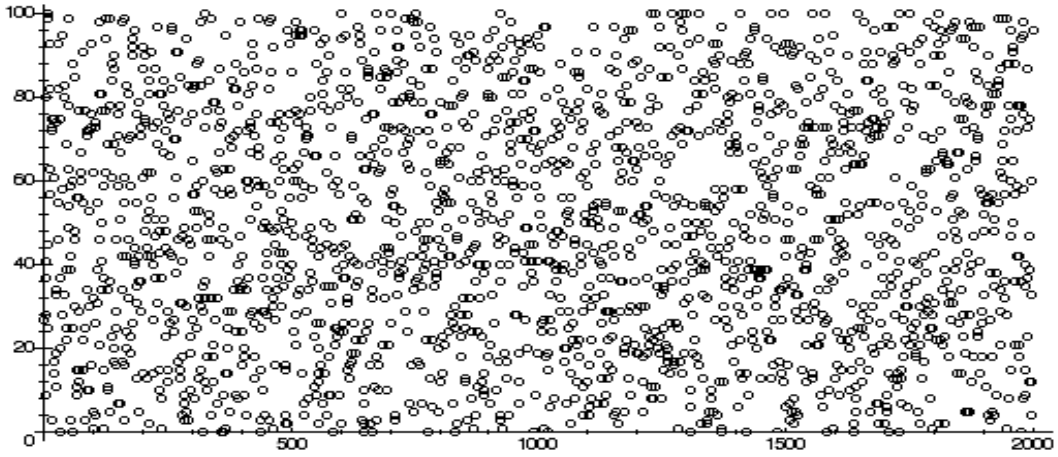


Figure 6: Scatterplot of two thousand uniformly distributed random integers between 0 and 100.

If however we waited long enough or if we compared completely unrelated languages, there is no reason to expect that the scatterplot of a random set of words (or an appropriate list such as the Swadesh list) would show any pattern at all. In fact we can create such random patterns using a very simple Maple command: *pointplot({seq([i,rand() mod 101], i=0..2000)});*

Now, we can see that we could possibly expect the $d_s d_p$ plots to fall along the diagonal because it is highly unlikely that the changes in meaning and phonological shape of sounds is going to be uncorrelated. As time passes, both the meaning and soundshape will tend to change together. Therefore instead of the scatter plot as in Figure (5) we will obtain as a correlated version as shown in Figure (7).

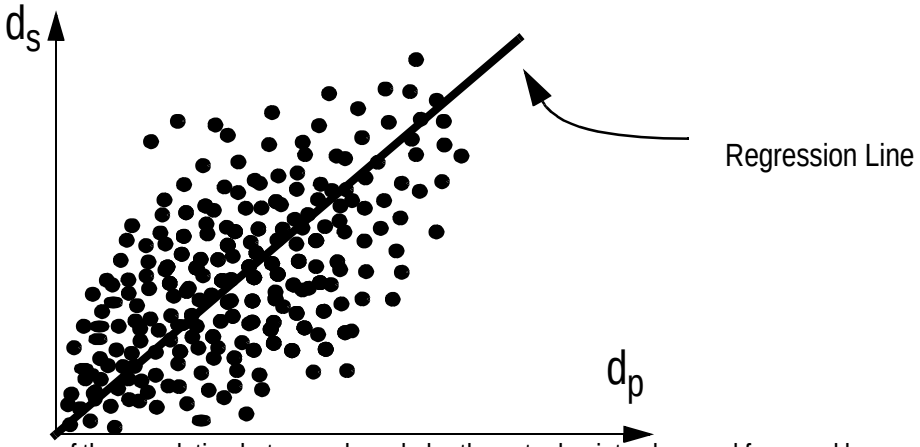


Figure 7 : Because of the correlation between d_p and d_s, the actual points observed from real languages might be skewed along a diagonal line (not necessarily 45 degrees). The line represents a hypothetical regression line.

If we observed absolutely no deterministic trend in the $d_s d_p$ scatterplots we would be justified in assuming that no kind of determinable relationship exists between the languages in question. In other words, we can *falsify* the presumed geneticity but we cannot really "prove" it in the sense of mathematics. We could however, have reasons for strong belief that there is determinism without knowing whether it is due to copying (borrowing) or due to descent from a common ancestor. The belief in descent from a common ancestor requires some assumptions, such as creating a Swadesh-like list and also assuming that some words such as ata, ana, apa, etc are infant-talk and are not genetic. This last form of reasoning is highly unlikely and should be dropped from methods of historical linguistics since there is plenty of evidence that the cart is being put in front of the horse [Hauser,1997], USNWR[1998], etc. It should be noted that the random walk and others like it use displacement in space as variables whereas the variables used in plots in Figures 5 and 7 are distances. Distances are always positive whereas displacements have direction (i.e. can have negative values). The reason for this seeming discrepancy is that we do not yet have a phonological or semantic space in which direction is defined. However, it is possible to work with metric spaces in which both negative and positive displacements are possible, as can be seen in Hubey[1994]. In this case negative displacements for semantics is still missing. These directions may be obtained via universals, by using the direction of the universal in the same way that the growth of entropy points in the direction of time, so that we may call entropy the arrow of time. In this case we would be looking at only a one-sided Gaussian function instead of the ones that were derived in the equations above. Therefore we would expect plots of the type shown in Figure (8). This makes the fact that data from some languages that seem to be distributed according to the Binomial Distribution (Ringe[1992]) completely irrelevant since it could have just as easily arisen from the random loss of lexemes of daughter languages as shown in the preceding mathematical models. A test similar to the Oswalt[Salmons&Joseph,1999] test in principle can be performed on the data in any of the plots in Figures (5) or (7). For example, if we permuted the words in the list, then the phonological and semantic distances in, say Figure (5) or (7) would no longer cause all the words in the list pile up at the origin or along the diagonal or to have the form of diffusion from the center as expected. Instead random phonological and semantic distances would result in plots resembling that of Figure (6).

.

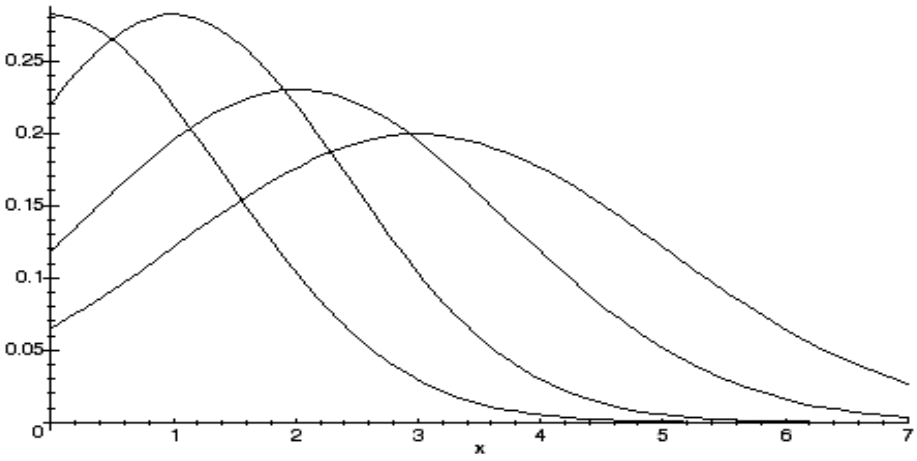


Figure 8: One-sided Gaussian Densities with various means and variances: using intuitive distance metrics could easily result in data resembling these one-sided Gaussian densities. It was shown earlier how these are very similar to the Binomial density used by Ringe to claim that data that fits such a density is an indication of nongeneticity.

6. Lexeme Loss, Glottochronology and Stochastic Differential Equations

It is known that the assumptions of glottochronology and lexicostatistics assumes a constant rate of lexeme loss. This is quite easily expressible as the differential equation

6.1) $$\frac{\dot{z}}{z} = -\gamma$$

It can be written as

6.2) $$\int \frac{dz}{z} = -\gamma \int dt$$

After integration we obtain

6.3) $$\ln(z) = -\gamma t + C \qquad or \qquad z = Ce^{-\gamma t}$$

Using the fact that both languages are changing at this rate, we can solve for the time of separation

6.4) $$t = \frac{\ln(z)}{2\ln(r)}$$

A detailed explication of this result has been done by Embleton [Salmons&Joseph,1998]. The final form can be found in other books, for example, Crowley[1992:181]. But this constant rate of change is the simplest approximation that can be made. We can change the constant γ to a function of time $\gamma(t)$ and add randomness w(t) (white noise) to make it more realistic as discussed above to obtain

6.5) $$\frac{dz}{dt} = -\gamma(t)z(t) + w(t)$$

But this is nothing but eq. (5.13) an equation which leads to the Gaussian probability density as shown. This means that randomness is expected in the word-list if because of natural change in languages so that finding randomness as claimed by Ringe does not imply that the word-list

achieved that particular state due to pure chance and having nothing to do with geneticity. In fact, the first order differential equation is solvable in all its generality, and the process of radioactive decay which is used in carbon-dating in a field allied with linguistics obeys the same kind of constant rate of loss, and the complete solution is given in Appendix F. Therefore with the assumption that the semantic and phonological changes in daughter languages are correlated we would need to simply look at scatterplots of semantic and phonological distances between daughter languages. To be more exacting, we should use correlation-regression analysis on the data. The catch here is that we need to be able to compute both semantic and phonological distances. Phonological distances can be easily computed. A simple way to accomplish this task in shown in Apendix D. More exacting and sophisticated distance metrics can be found in Hubey[1994]. However what we really need are semantic distances. In order to do this we need a more thorough comprehension of measurement theory. A simple discussion of the topics of measurement scales can be found in Appendix E, and more can be found in Hubey[1998b], and also Stevens [1966] or Torgerson [1958]. There are additional problems with the formulation (6.2); one big problem is that the randomness as above is Gaussian and that is not really appropriate since we cannot gain what was lost so that we should really use a one-sided noise density. This can be accomplished still within the framework of stochastic differential equations as above, but it does not fit into the scope of the present work. One can continue and compute distances between languages. Obviously, such distances are of the form d=vt where v=rate of change and t=time. Since little is known about rates of change in languages or effects of contact we could try a more general form

$$= \gamma C^{\alpha} t^{\beta} \tag{6.6}$$

where we have used C, the intensity of contact of the language with other languages, as a proxy for 'rate of change'. This equation or others like it may be interpreted very loosely as "the distance between languages (which is a function of number of words still in use which were derived from a common ancestral language) is a function of both the passage of time and the rate of loss" and therefore the multipication above can be interpreted as a kind of fuzzy logical-AND. If we assume that the rate of loss is proportional to "intensity of contact" with other languages we can create other mathematical models for the interaction of language families with relatives and non-related languages. All of this is beyond the scope of this work. However, we can make some general comments. In Figures (5) and (7) we imagined that we were tracing the diffusion of the same set of words in daughter languages accross time. Those figures are driven by both phonology and semantics. However, in general if we are to compare words from any two languages, we can make the scatterplots either as semantically-driven or phonologically-driven. In the case of the phonologically driven scatterplots, we would simply start with some basic-word list and then find words in the other language which phonologically matched these words. If we plotted them, we would find that the phonological distances would be small, but there would be no correlation in semantics so that it would be probably uniformly (randomly) distributed. Therefore we would expect to see a scatterplot as shown in Figure (9a) or (9b). The very fact that we might obtain a scatterplot similar to (9b) itself should make us wonder why two unrelated systems should wind up with such similar phonological and morphophonemic systems. Is it a "universal" (whatever it means) or is there some long-range genetic relationship. After all, the lexicon can conceivably get completely completely changed over time, but if the intensity of contact is low, there is no reason why such system characteristics as morphology, morphophonemics, and phonological systems should converge. In any case, nothing like this is done in linguistics. Instead the scatterplot is semantically-driven in that the words matching in meaning (or with similar meanings) are compared so that we

really look for a phonological match for corresponding semantic matches as in Figure (9c). In that case, there is no reason to expect any phonological similarities (i.e. small distances) so that we should obtain something like Figure (9c). In no case, should we obtain anything like in Figure (5) or Figure (7). Those clearly fall into the range of normal random diffusion of meaning and phonology of the lexicon of a language over time. To be precise, it should be possible to compute some statistics for every pair of languages using methods such as correlation-regression analysis, either linear or those appropriate for the diffusion equation or the heat equation. These statistics can then be compared for any pair of languages. It is not necessary for every linguist to know every language; it is not necessary to argue from authority, and it is not necessary to be dependent on other linguists' gut feelings.

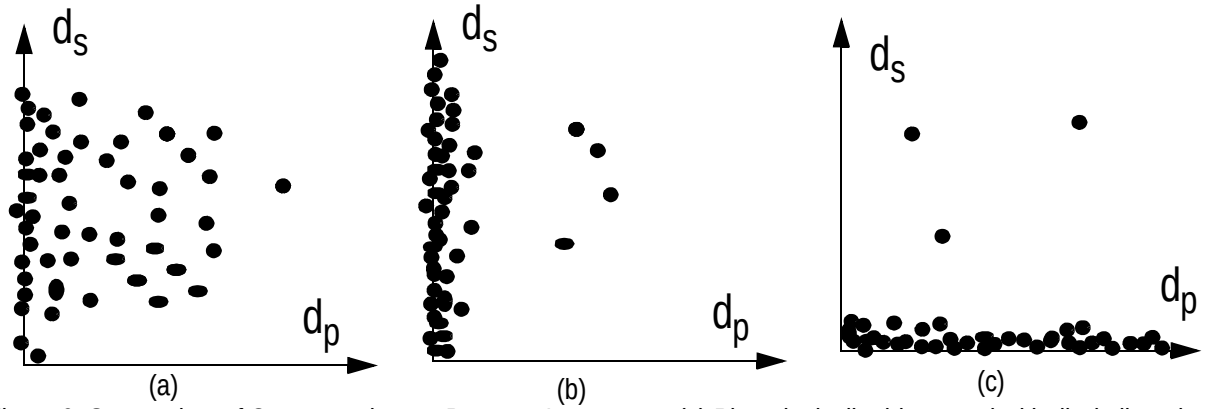


Figure 9: Scatterplots of Correspondences Between Languages: (a) Phonologically-driven, and with dissimilar phonological systems; (b) Phonologically-driven, and with similar phonological and morphophonemic systems; (c) Semantically-driven scatterplots of correspondences between competely unrelated languages. Traditional linguists make a real big issue out of the fact that some small numbers of points do not all along the abcissa in (c) if statistical methods are used [see for example, Trask[1996]). The simple fact that the assumption of Gaussian errors is used in Labov's work also, which he extols, is apparently not understood.

Conclusion:

The equations, the figures, plots and the narrative are self-explanatory. It should be noted that what has been shown is partially answers to vexing questions and partially a program of research for the future using the methods and mathematical models shown and explained above. The solution to the chance cognacy problem is given in Appendix A. The standard heuristic method, called the comparative method, is a heuristic, nothing more, and not clearly understood by even those who write books. There are many wrong answers, but only one right answer in science. The methods presented here are precise and to various degrees of accuracy, and should be improved with time after experience using them. The methods such as Ringe's and Oswalt's are probably best tested using simulation. It is under such circumstances that we really are 100% certain of the answers. Indeed the equations and solutions offered here should also be tested via simulation. The results of Appendix A (that on average only a single word would be found by chance to match between unrelated languages) has already been tested via simulation. Oswalt's heuristic can be understood in terms of the correlation functions of mathematics and physics, and similar ideas can be used to modify Ringe's "method" more meaningful but it is beyond the scope of the present paper.

Appendix A: Solution of the False Cognacy Problem

The problem may be stated as: Given two languages, what is the average number of words that they will have in common with the same meaning purely due to chance and not due to borrowing or common ancestry? This formulation of the problem avoids the problem of having to posit particular phoneme systems for languages in question and then create CV, VC, CVC, CCVCC etc. syllables and compute their probabilities.

Let's first denote one language as a kind of a standard S against which we will compare the others. We then calculate how many languages can exist such that no word matches any word in this standard language; and call this number Z(N) and call these languages L. The one standard language, S, is the language whose words are ordered in some fixed way, say 123.....N. Let's also call L_i the set of languages whose ith semanteme corresponds to word i (i.e. the ith word of this language matches the ith word of the S language). So $L_{i,j}$ is then the language whose ith and jth words match. For example for N=4,

A.1)
$$L_2 = \{1234,1243,3214,3241,4213,4231\}$$
$$L_{2,3} = \{1234, 4231\}$$
$$L_{2,3,4} = \{1234\}$$

Thus L_{ij} contains all languages whose ith and jth words match. From the definition, it follows that

A.2) $L_i \cap L_j = L_{ij}$ $L_{ij} \cap L_{klm} = L_{ijklm}$ where $i \neq j; i \neq k; i \neq l; etc$

We can then compute the number of elements of each set using $|L| = N!$ which is the number of possible languages (one of which is the S language) $|L_i| = (N-1)!$ since only the ith word will match and the rest can be distributed as above. Therefore $|L_{ij}| = (N-2)!$ using the same reasoning as above. In general $|L_{n_1 n_2 \ldots n_j}| = (N-j)!$ and $|L_{ij\ldots N}| = 1$. Let Q be the set of languages which have zero match. We can write Q in terms of the subsets that are defined above. Any language which has at least one match is a member of at least one L_i for some appropriate value of i. So any language which is not in the set $_1 \cup L_2 \cup \ldots \cup L_N$ has no match. Thus the number of elements in Q is:

A.3) $$Z(N) = |Q| = N! - |L_1 \cup L_2 \cup \ldots \cup L_N|$$

We can calculate the number of elements in $_1 \cup L_2 \cup \ldots \cup L_N$;

A.4) $$|L_1 \cup L_2 \cup \ldots \cup L_N| = |L_1| + |L_2| + \ldots + |L_N| + \quad + (-1)^{N-1} |L_{123\ldots N}|$$

This is an application of the inclusion-exclusion principle which can be found in any introductory book on set theory or discrete mathematics. There are N terms with magnitude (N-1)!, then the negative terms follow with magnitude (N-2)! and their number is N(N-1)/2= C(N,2), etc where the C(N,k)= N! /(k!(N-k)!). So we have

A.5) $$\left|L_1 \cup L_2 \cup \ldots \cup L_N\right| = C(N,1)(N-1)! - C(N,2)(N-2!) + (-1)^{N-1}C(N,N)(N-N!)$$

$$\left|L_1 \cup L_2 \cup \ldots \cup L_N\right| = \frac{N!}{1} - \frac{N!}{2} + \frac{N!}{3} - \ldots + (-1)^{N-1}\frac{N!}{N!}$$

We can get an approximate result from this as;

A.6) Z(N)= N!/e where = 2.171 the the base of natural logarithms.

So we can calculate the number of languages which have exactly k matches. The number of possible k pairs is C(N,k). We want these k words' semantemes to match but we don't want the N-k semantemes to match. The number of ways of doing this is Z(N-k). Since the number of possible languages with exactly k matches with language S, is given by C(N,k) Z(N-k) then the probability of finding exactly k matches is the ratio of the numbers. If we want to see how easy comparisons have been made because of messy matching we have to use the inclusion-exclusion principle to factor out duplications.

Now for another solution, let M_j be the event that the jth word is matched regardless of what happens to the other words. If the language has N words then there are N! permutations of meanings and words. If the jth words match then that leaves (N-1)! permutations for the rest of the words so that

A.7) $$P(M_j) = \frac{(N-1)!}{N!} = \frac{1}{N}$$

Similarly if the jth and kth words match, where $j \neq k$, that leaves (N-2)! permutations for the remaining words so

A.8) $$P(M_j M_k) = \frac{(N-2)!}{N!} = \frac{1}{N(N-1)}$$

and for three

A.9) $$P(M_j M_k M_l) = \frac{(N-3)!}{N!} = \frac{1}{N(N-1)(N-2)}$$

etc. The probability we are looking for is

A.10) $$P(M_{k_1} M_{k_2} \ldots M_{k_m}) = \frac{(N-M)!}{N!} = \frac{1}{(N-M+1)(N-M+2)\ldots(N-1)(N)}$$

For large N and small M this is approximately $1/N^M$. For N=100 and M=3 the number is $100^{-3} = 10^{-6} = 0.000001$. Of course, this is not the probability that m words in each of two languages with n words can be found to match each other. This is the probability that if we selected 3 words at random from each of two languages with 100 words each, that all three would match. To find the probability of at least one match we must add all of these but we have to account for all possible ways of matching using the inclusion-exclusion principle [see for example Hubey[1998b]. Therefore we must add

A.11) $$C(N,1)\frac{1}{N} - C(N,2)\frac{1}{N(N-1)} + C(N,3)\frac{1}{N(N-1)(N-2)} - \ldots + (-1)^{N-1}\frac{1}{N!}$$

which is

A.12) $$1 - \frac{1}{2!} + \frac{1}{3!} - \quad + \ldots + (-1)^{N-1}\frac{1}{N!}$$

We know from Taylor Series expansions that

A.13) $$1 - e^{-1} = 1 - \frac{1}{2!} + \frac{1}{3!} - \quad + \ldots + (-1)^{N-1}\frac{1}{N!} + \ldots = \sum_{N=1}^{\infty} (-1)^{N-1}\frac{1}{N!}$$

Since this series converges very rapidly 1-e^{-1} can be used as a very good approximation. The expected number of matches can be calculated rather easily from the definition of the mean:

A.14) $$\langle N \rangle = \sum_{j=1}^{N} P(M_j) = N \cdot \frac{1}{N} = 1$$

This problem has already been solved in another context and can be found, for example, in Graham, Knuth and Patashnik [Graham et al, 1989] or Roberts [1984]as the 'hat-matching' problem. We can consider this to be the 'lexeme-matching problem' The development here follows the book. Suppose there are only four lexemes (N=4) and the lexemes are named A, B, C and D, then there are N!=24 ways for the semantemes to land in their lexeme containers as shown in the table below. Since the size of the problem is small an exhaustive enumeration can be used to show the possibilities. The number of matches is shown next to the permutation:

Table 1: Some Permutations and Their Number

ABCD=4	ABCD=4	CABD=1	CABD=1
ABDC=2	BADC=0	CADB=0	DACB=1
ACBD=2	BCAD=1	CBAD=2	DBAC=1
ACDB=1	BCDA=0	CBDA=1	DBCA=2
ADBC=1	BDAC=0	CDAB=0	DCAB=0
ADCB=2	BDCA=1	CDBA=0	DCBA=0

Thus h(4,4)=1; h(4,3)=0; h(4,2)=6; h(4,1)=8; h(4,0)=9. In general, there's only one way in which every semanteme will fall in its own lexeme (i.e. N matches); there is no way in which there can be N-1 matches and the largest number is zero matches. We can gain further insights into this problem and its solutions by examining the solution of the general problem. "A group of N fans of the winning football team throw their hats high into the air. The hats come back randomly, one hat to each of the N fans. How many ways h(N,k) are there for exactly k fans to get their own hats back?" This is really the same problem we just solved above. "We can determine h(N,k) by noticing that it is the number of ways to choose k lucky hat owners, namely C(N,k), times the number of ways to arrange the remaining N-k hats so that none of them goes to the right owner, namely h(N-k,0). A permutation ,... and we have the general formula" [Graham et al, 1989]

A.15) $$h(N,k) = C(N,k)h(N-k,0)$$

Some of the computed terms from Graham et al [p.194] are

Table 2:

N	h(N,0)	h(N,1)	h(N,2)	h(N,3)	h(N,4)	h(N,5)	h(N,6)
0	1						
1	0	1					
2	1	0	1				
3	2	3	0	1			
4	9	8	6	0	1		
5	44	45	20	10	0	1	
6	265	264	135	40	15	0	1

The same result as before, namely the probability of no fans getting their hats back (or a semanteme falling on its lexeme) is computed to be 1/e. The average number of matches turns out to be *exactly one*; that is, on the average the two languages will only have one word in common which has the *same meanin*g. Furthermore, we can compute the various probabilities: P(0)=265/720=0.3681, P(1)=264/720=0.3667, P(2)=135/720=0.1875, P(3)=40/720=0.556, P(4)=15/720=0.208 so that these numbers are practically identical to the Poisson approximation as shown above (and to the binomial numbers derived by Ringe). Indeed, even for N=5, we obtain P(0)=44/120=0.3667, P(1)=0.3750, P(2)=0.1667 so that there is rapid convergence.

The average number of words that will match in two unrelated languages is *one*, only a single one, if both languages have exactly the same phonemes, and exactly the same phonological words. So if in real languages, each word, on average has about 3 meanings and if we allow some loose matching of sounds up to 5 others (that sound something like each other) we could then have 3*5 or 15 matches. If we allow for 10 semantic shifts and say 5 sound-alikes we'd get about 50 matches due to chance. As can be easily seen, that is already pushing the limit. We could probably put some top limit on accidental matches of around 30-50 depending on what kinds of things are being counted as matches (which depends on the people doing this, obviously).

The complete solution is now obvious. If we allow messy-matching (which is someting linguists always complain about) we only need to multiply the solutions obtained to factor in the problems introduced by such messy-matching. If we allow a phonological shift/factor of P, and with semantic shift/factor of S, then the average number of words that could match due chance in unrelated languages is simply P*S. For example for a P=10, and S=5 we would expect about 50 matches due to chance. Some of the numbers derived by others alleging that 400 accidental matches can be expected seem to be utterly absurd. In any case we can account for messy-matching in a variety of ways. For example, sometimes amateurs in linguistics (or protoworlders or Nostraticists according to some) try to match words from different languages simply by finding sounds that are simi-

lar. By this word we obviously mean something that we think of as the opposite of or fuzzy complement of distance which measures difference. What the amateurs do is really like what the professionals do in that they group phonemes into sets which are similar to each other. For example, the phoneme /f/ might sometimes be grouped with /p/ (because it is labial) or with /th/ (i.e. t because it is a front fricative) or with /s/ or /sh/ (because it is still a fricative).

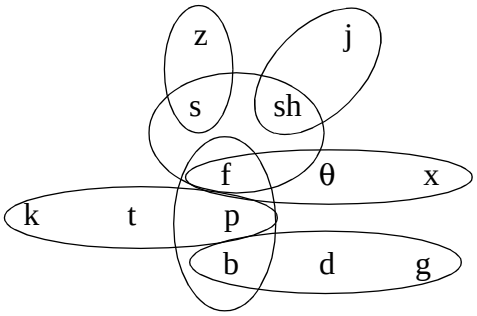


Figure A.1: Messy matching: averaging averages

For each sound count the number of matches possible and then average them:
z=2 ; s=4; j= 2; sh=4; f=7; θ=3; x=3;
k=3; t=3; p=6; b=3; d=3; g= 3;

The average is 46/13=3.538 Allowing for the sounds in the sets in the Venn diagrams to be counted as matches we would get approximately 3.5 times as many matches as if we only used exact matchings (if the distributions of these sounds were uniform).

To get rigorous results using this “messy matching” we have only to use the concept of distance (appendix D, and also Hubey[1994]) and then use probability methods such as correlation-regression analysis or try to fit data into mathematical models.

Appendix B: Examples of Uses of Bayesian Reasoning

Another simple example of type of calculation that might be used in historical linguistics is this: a protolanguage some vowels and consonants. The perceptual mechanism and the sound creation mechanism is such that 2/5 of the vowels and 1/3 of the consonants get mistaken for the other. The ratio of the vowels to consonants in the language is 5 to 3. What is the probability that what is heard is what is sent if (i) the received sound is a vowel, (ii) the received sound is a consonant.

Let A be the event that a vowel is heard and B that a consonant is heard. Let H_v be the hypothesis that the signal was a vowel and H_c be the hypothesis that the signal was a consonant. By the phonotactics of the language $P(H_v)/P(H_c)=5/3$. We also know that $P(H_v)+P(H_c)=1$. Hence $P(H_v)=5/8$ and $P(H_c)=3/8$. We also know that $P(A|H_v) = 3/5$, $P(A|H_c) = 1/3$, $P(B|H_v) = 2/5$, and $P(B|H_c) = 2/3$. We compute the probabilities of A and B from the total probability formula.

B.1) $$P(A) = \frac{5}{8}\cdot\frac{3}{5}+\frac{3}{8}\cdot\frac{1}{3} = \frac{1}{2}, \text{ and } P(B) = \frac{5}{8}\cdot\frac{2}{5}+\frac{3}{8}\cdot\frac{2}{3} = \frac{1}{2}.$$

From these we compute

B.2) $$P(H_v|A) = \frac{P(H_v)P(A|H_v)}{P(A)} = \frac{\frac{5}{8}\cdot\frac{3}{5}}{\frac{1}{2}} = \frac{3}{4}$$

B.3) $$P(H_c|B) = \frac{P(H_c)P(B|H_c)}{P(B)} = \frac{\frac{3}{8}\cdot\frac{2}{3}}{\frac{1}{2}} = \frac{1}{2}$$

As an example possibly more closely related to processes of historical linguistics, suppose a linguist is in the process of deciding if a given word is descended from language L_1, L_2 or L_3. He has determined that these probabilities are 0.5, 0.3 and 0.2 from the overall word structure of the vocabulary. Suddenly some new materials are found by archaeologists which indicate some information I which it is thought will lend support to one of these languages. The linguist after some more thought decides that the probabilities of this information lending support to L_1, L_2, or L_3 are 0.6, 0.2, and 0.1 respectively. How should he revise his probabilities depending on this new evidence?

B.4) $$P(L_k|I) = \frac{P(I|L_k)P(L_k)}{P(I|L_1)P(L_1)+P(I|L_2)P(L_2)+P(I|L_3)P(L_3)}$$

The denominator is 0.38 so $P(L_1|I) = 0.79$ $P(L_2|I) = 0.16$ $P(L_3|I) = 0.05$.

Appendix C: Some Probability Mass Functions and Densities

Geometric Distribution. An experiment is performed in which only the occurrence or non-occurrence of an event is recorded. The probability of success (occurrence) is p and the probability of failure (non-occurrence) is q = 1 - p. The successive trials of the experiment, then, are known as independent Bernoulli variables. The mass function of the rv x, which gives the number of trials until a success is recorded is given by the Geometric Distribution:

C.1) $$f(x) = p(1-p)^{x-1}$$

Binomial Distribution. The Binomial Distribution gives the probability of obtaining x successes in n trials in a sequence of Independent Bernaulli distributed random variables. The Binomial Distribution can be derived as follows. One of the ways in which we can have x successes and n - x failures is to have x successive successes followed by n - x successive failures. This probability

C.2) $$\underbrace{pp\ldots p}_{x}\underbrace{qq\ldots q}_{n-x} = p^x q^{n-x}$$

However, the successes and failures may be combined in any order to give x successes, so that using the number of permutations of such successes we obtain the Binomial Mass function

C.3) $$f(x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}$$

Poisson Distribution. Consider the occurrence of an event such that the "occurrence" or "arrival" of events in a time interval or space interval t is independent of earlier intervals, and is independent of the time t then the random variable x is Poisson distributed. The density function is given by

C.4) $$f(x) = \frac{e^{-\lambda\tau}(\lambda\tau)^x}{x!}$$

where λ="arrival" rate and τ=time interval such that μ=λτ=mean of the mass function. In the limit that $p \to 0$ and $n \to \infty$ such that $np \to \mu$ the binomial mass function approaches the Poisson Distribution with mean μ. Hence, the Poisson mass function serves as a good approximation to the binomial for large n and small p.

Appendix D: Distance and Phonetic Similarity

Phonemes as Fuzzy Vectors

We can have low resolution phoneme description using only five dimensional fuzzy vectors using the dimensions front-back, round-unround, nasal-nonnasal, voiced-unvoiced and motion-steady_state.

D.1) $$\mathbf{u}=\begin{bmatrix}0 & back\\ 1 & round\\ 0 & non-nasal\\ 1 & voiced\\ 0 & no-motion\end{bmatrix} \quad \mathbf{\ddot{u}}=\begin{bmatrix}1 & front\\ 0.2 & round\\ 0 & non\text{–}nasal\\ 1 & voiced\\ 0 & no-motion\end{bmatrix} \quad \mathbf{f}=\begin{bmatrix}x & dont\text{care}\\ 0.1 & labio-dental\\ 0 & not-nasal\\ 0 & voiceless\\ 0 & no-motion\end{bmatrix}$$

D.2) $$\mathbf{v}=\begin{bmatrix}x & dont\text{care}\\ 0.1 & labio-dental\\ 0 & not-nasal\\ 1 & voiced\\ 0 & no-motion\end{bmatrix} \quad \mathbf{m}=\begin{bmatrix}x & dont-care\\ 0 & closed\\ 1 & nasal\\ 1 & voiced\\ 0 & no-motion\end{bmatrix} \quad \mathbf{n}=\begin{bmatrix}0.5 & half-open\\ 0.5 & round\\ 1 & nasal\\ 1 & voiced\\ 0 & no-motion\end{bmatrix}$$

D.3) $$\mathbf{p}=\begin{bmatrix}x & dont-care\\ 0.1 & open-close\\ 0 & non\text{–}nasal\\ 0 & voiceless\\ 1 & motion\end{bmatrix} \quad \mathbf{b}=\begin{bmatrix}x & dont-care\\ 0.1 & open-close\\ 0 & non\text{–}nasal\\ 1 & voiced\\ 0.5 & motion\end{bmatrix} \quad \mathbf{k}=\begin{bmatrix}0.5 & neutral\\ 0.1 & open-close\\ 0 & non\text{–}nasal\\ 1 & voiceless\\ 1 & motion\end{bmatrix}$$

If we had distances for semantics (maybe in some localized space) we could then use these phonological distances to regress against semantics.

Binary Features

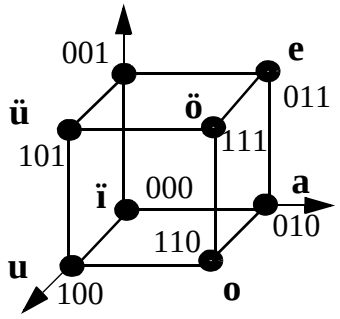
Generally, the distinctive features of phonology are binary. See for example, Clark & Yallop [1990], or Lass [1984]

D.4) $$\mathbf{p}=\begin{bmatrix}+back\\ +nasal\\ \dots\\ \text{–}voicing\end{bmatrix} \text{ or as } \mathbf{p}=\begin{bmatrix}1\\ 0\\ \dots\\ 1\end{bmatrix}$$

For space savings we might write it as p_i where the n-tuple is assumed to have values in {0,1} instead of the +, or - that are commonly used which we can represent simply as a bitstring such as 10001010...1011. In this case the Hamming distance between phonemes is nothing more than the bitwise XOR of the two bitstrings representing the phonemes so that

D.5) $$d(p,q) = p_1 \oplus q_1 + p_2 \oplus q_2 + \ldots + p_n \oplus q_n = \sum_{n} p_i \oplus q_i$$

This distance will work for anyone's binary distinctive features. How well these distances correlate with perceptual distances is an empirical question and probably will require more sophisticated distance metrics [Hubey,1994]. However, a very simple distance metric is given below. We can compute distances between phonemes using different ways.



Hamming Distance example

$$d(i,o) = i_1 \oplus o_1 + i_2 \oplus o_2 + i_3 \oplus o_3 = 3$$

Euclidean Distance example

$$d(i,o) = \sqrt{(|0-1|)^2 + (|0-1|)^2 + (|1-0|)^2} = \sqrt{3}$$

Figure D.1: Distance Examples on the Ordinal Cube of Wowels [Hubey,1994]

The 3D vowel cube can be seen to be many things in one. We can think of it as a vector space for vowels except that it has to be twisted and sheared into shape to fit the actual human speech sounds based on formants [Hubey,1994]. One needs to use the 3D versions of the scale, shear, translate and rotate matrices to fit the cube into data obtained from actual speech of humans.

Appendix E: Measurement Scales

Before we try to measure or normalize quantities we should know what kinds of measurements we have. They determine if we can multiply those numbers, add them, rank them etc. Accordingly measurements are classified as: (i) Ratio scale, (ii) Interval scale, (iii) Ordinal scale, or (iv) Nominal scale.

Ratio Scale: The highest level of measurement scale is that of ratio scale. A ratio scale requires an absolute or nonarbitrary zero, and on such a scale we can multiply (and divide) numbers knowing that the result is meaningful. The standard length measurement using a ruler is an absolute or ratio scale.

Distance Probably the most common measurement that people are familiar with is that of distance. It is such a general and common-sensical idea that mathematicians have abstracted from it whatever properties it has that makes it so useful and have extended it to mathematical spaces so that this idea, is in fact, used and useful in the previous ideas of measurements. The requirement that the concept of distance satisfies is this:

E.1) $$d(x, z) \leq d(x, y) + d(y, z)$$

The concept of "distance" or "distance metric" or "metric spaces" is motivated by the simple concept illustrated below.

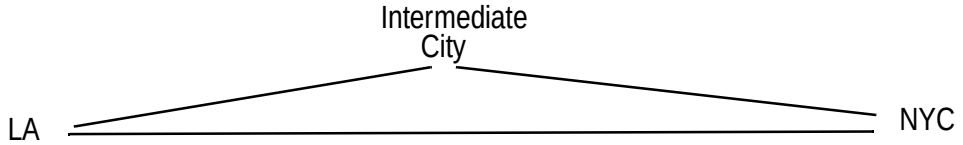


Figure E.1: Concept of Distance Metric: Any detour from NYC to LA cannot be a shorter than the direct distance between the two since distance is measured as the shortest distance between two points.

If we substitute from the figure above we can see that the distance from LA to NYC can never be greater than the distance from LA to some intermediate city plus the distance from that intermediate city to NYC. Any space in which distance is defined is a metric space.

Hamming distance is the number of bits by which two bitstrings differ. For example the distance between the bitstring 1111 and 0000 is 4 since the corresponding bits of the two bitstrings differ in 4 places. The distance between 1010 and 1111 is two, and the distance between 1010 and 0000 is also two. In phonology, the basic primitive objects are phonemes. They are descriptions of the basic building blocks of speech and are usually described in binary as the presence or absences of specific characteristics such as voicing, rounding, frication, plosivity etc. Since we can represent these as bitstrings the Hamming distance can be used to measure the distance between phonemes [Hubey, 1994] and also in Appendix D. Such hypercubes can be constructed also for sets of consonants, as seen below and also in Hubey[1994]. We need higher dimensional or fractional spaces to represent them all on the same graph or hypercube or same space

.

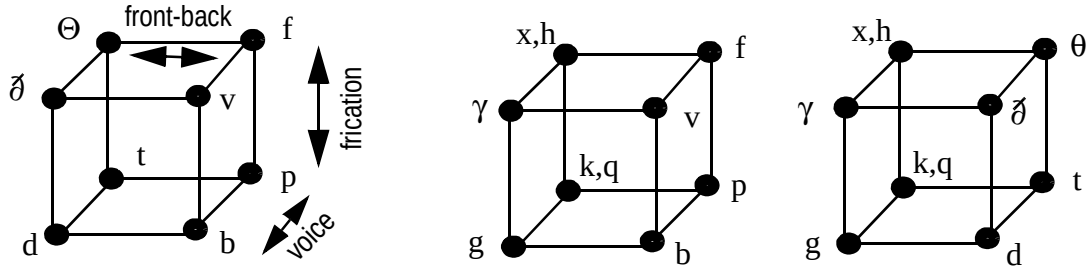


Figure E:2: Low Resolution Representation of Some Consonantal Phonemes: The three basic dimensions are voicing, frication, and place of major constriction.

In order to represent them all there are several ways. One way is to use a ternary system in which the digits {0,1,2} are used in some positions as shown below:

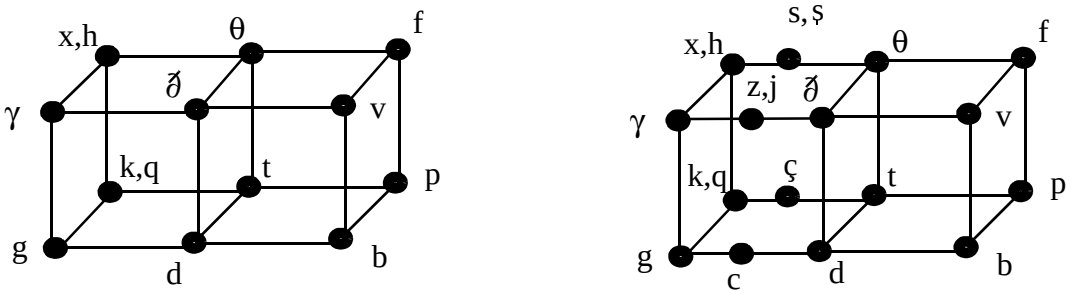


Figure E:3: Low Resolution Representation of Some Consonantal Phonemes

We can add to the ternary system by making it fractional as shown above on the right or make it quaternary as shown below:

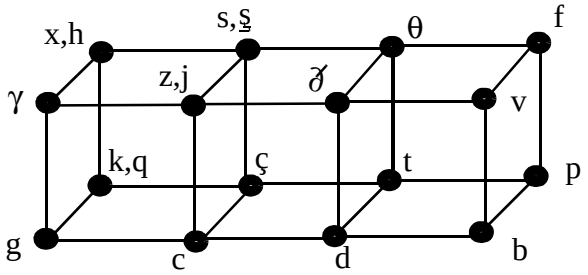


Figure E:4: Low Resolution Representation of Some Consonantal Phonemes

We still do not have such common phonemes as /m/, /n/, /r/, or /l/ in the above figures and more sophisticated spaces are needed [Hubey,1994]. Whether these distances are directly proportional to perception is an empirical question, and the correlation between the perceptual and acoustic values can be found in experiments conducted for decades by many phoneticians, see for example the works of Lindblom, Stevens, Blumstein, and co-workers [Liberman & Blumstein,1988].

Interval Scale: However, not everything that can be measured or represented with integers (or real numbers) is not a ratio/absolute scale. For example, the Fahrenheit temperature scale is only an interval scale. The differences on an interval scale (such as the Fahrenheit scale) are valid and meaningful and correct, but multiplication is not. For example, 100F is not twice as hot as 50F.

Ordinal Scale: The next level on the measurement scale is the ordinal scale, a scale in which things can simply be ranked according to some number but the differences are not valid. In the ordinal scale we can make judgements such as A>B. Therefore if A>B and B>C, then we can conclude that A>C. In the ordinal scale there is no information about the magnitude of the differences between elements. We cannot use operations such as +, -, * or / on the ordinal scale. It is possible to obtain an ordinal scale from questionnaires. One of the most common, if not the most common is the multiple-choice test which has the choices: extremely likely/agreeable, likely/agreeable, neutral, unlikely/disagreeable, and extremely/very unlikely/disagreeable. This is the common Likert Scale used in polls.

Nominal Scale: The lowest level of measurement and the simplest in science is that of classification. In classifying we attempt to sort elements into categories with respect to a particular attribute. This is the nominal scale. On this scale we can only say if some element possesses a particular attribute but cannot even rank them according to some scale on a hierarchy based on the intensity of possession of that attribute. We can only think of creating sets based on the possession of some property and apply the operations for sets. In this sense the set operations are the most primitive of operations of mathematics. It ranks so low on the scale or hierarchy that we all instinctively do it. Whatever kind of logic that flows from this must obviously be related to set theory in some way.

Accuracy and Precision

There is usually no thought given to the possibility of measuring something accurately but not precisely, or precisely but not accurately. Suppose we want to compute the area of a rectangle with width w and length L, but the measurements are not and can never be to infinite precision but have errors in them as shown. Therefore our calculation is really $A = (L \pm \Delta L)(W \pm \Delta w)$ instead of A=Lw where DL and Dw are the errors with which we measure L and W.

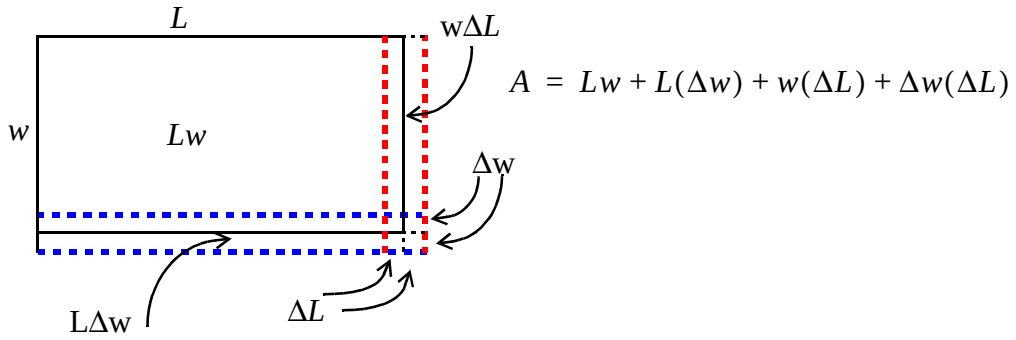


Figure E:5: Errors in Measurement: By choosing A=Lw we mean that we drop the other small terms which are small error terms that we cannot get rid of.

Therefore $A = Lw + L(\Delta w) + w(\Delta L) + \Delta w(\Delta L)$. If the error is about one tenth of the actual size that is measured, then the last product is about one-hundredth and can be dropped introducing no more than an error of one-hundredth in the final answer for the area. Dropping the other terms will introduce an error of about ten percent. As an example, suppose we measured L=5.2 and w=3.1232123432. The product is 16.2407041846 but all of these digits are not significant in the

sense that they are a part of the error since L has not been measured to any more accuracy than 2 digits. For example the error in L can be on the order of 0.1, therefore L could in reality be between 5.0 and 5.2. If we use these then we see that 15.616061716 < A <16.553025419. Therefore since the lowest precision number is L, and since it only has 2 significant digits (i.e. the 5.2 since the last digit could be error) then the answer A is only significant to two digits; therefore A=16.0 (correct basically to two significant digits) and which in a sense is an average since it is something like the median value in the interval [15.6,16.5]. This number is approximately equal to the average of the upper and lower limits of A whose first few digits are is 16.08. In simple terms when we make computations such as finding the area of a piece of land we should make sure that our calculated answers do not give the appearance of being better than they are. However the resolving power of our tools is and must always be greater than the reach of our concepts. It is easy to know the resolution of our physical instruments and thus perform error analysis. However, in the non-physical sciences we are forced to create different kinds of instruments with which to measure things. There are different types of problems associated with social sciences.

Reliability and Validity

So habituated are we to measuring things in this modern age that we scarcely give thought to the possibility that what is being represented as a number may be meaningless. That is the validity of the measurement i.e. that the measurement or metric actually measures what we intend to measure. In physical measurements there is usually no such problem. Validity also comes in different flavors such as construct-validity, criterion-related validity, and content-validity. Reliability refers to the consistency of measurements taken using the same method on the same subject.

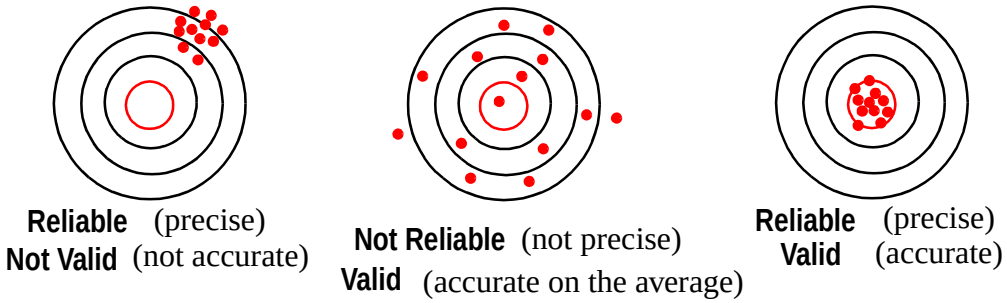


Figure E.6: Reliability and Validity Analogy: One normally expects accuracy to increase with precision. However in the social sciences they are independent.

In paleontology one is often required and forced to make deep claims on the basis of partial bones of long dead living things. In order to find some regularity one is forced to take into account basically shapes and size of bones from which conclusions are obtained about the species. Obviously, precision and accuracy are highly correlated in these eye-balling measurements. Judging from the tremendous variation in size and shape of a single living species such as dogs, one should be very careful announcing that a given set of bones belongs to a different species of hominids. There are two species of monkeys for which the differences in skeletons cannot be ascertained. There are other species (for example, dogs) whose sizes vary greatly. Therefore much of the *lumper vs. splitter* arguments are probably not decidable from the fossils since the precision of the instruments and theory is insufficient to make such determinations.

Appendix F: Differential Equations and Carbon Dating

Living cells absorb carbon directly or indirectly from carbon dioxide (CO_2) in the air. The carbon atoms in some of this CO_2 are composed of a radioactive form of carbon ^{14}C rather than the common ^{12}C. ^{14}C is produced by the collisions of cosmic rays (neutrons) with nitrogen in the atmosphere. The ^{14}C nuclei decay back to nitrogen atoms by emitting b-particles. All living things, or things that were once alive, contain radioactive carbon nuclei. In the late 1940s, Willard Libby showed how a careful measurement of the ^{14}C decay rate in a fragment of dead tissue can be used to determine the number of years since its death. In any living organism the ratio of ^{14}C to the total amount of carbon in the cells is the same as that in the air. If the ratio in the air is constant intime and location, then so is the ratio in living tissue. After the organism is dead, ingesion of CO_2 ceases and only the radioactive decay continues. The internationally agreed upon half-life t of ^{14}C is taken to be 5568 ∓ 30 years. Let q(t) be the amount of ^{14}C per gram of carbon at time t in the charcoal sample (where time is measured in years). Let t=0 be the present and T <0 be the time that the sample died. Then q(T)=q_T > 0 is a constant. For t>T the 14C nuclei decay is given by the first order differential equation

F.1) $$\frac{dq}{dt} = -kq \qquad q(0) = q_0 \qquad T \le t \le 0$$

The solution of the equation is is $q(t) = q_0 e^{-kt}$. The half-life t is defined to be the time in which the radioactive material is half of its initial value, that is $\frac{q(\tau)}{q_0} = e^{-k\tau} = \frac{1}{2}$ so that this equation is equivalent to $\ln(2) = k\tau$. In this case $k \approx 0.0001245$. Therefore at the time of death (t=T) the solution is written as $q_T = q_0 e^{-kT}$ so that solving for T we obtain

F.2) $$T = -\frac{1}{k}\ln\left(\frac{q_T}{q_0}\right)$$

We now need to calculate the ratio q_T/q_0. which is the ratio of the radioactive carbon to the nonradioactiave. However we do not have this information. We only know the rate of decay of the sample at the time we make the measuements. According to the ODE (ordinary diff. eq.) the rate of disintegration of the radioactive nuclei at time t is proportional to the amount present. Thus

F.3) $$\frac{q'(T)}{q'(0)} = \frac{kq(T)}{kq(0)} = \frac{q(T)}{q(0)}$$

Suppose that the sample decay rate is measured at 1.69 disintegrations per gram of carbon per minute, while in living tissue there were 13.5 disintegrations per gram of carbon per minute. Therefore we can see that the final solution is

F.4) $$T = -\frac{1}{k}\ln\left(\frac{q_T}{q_0}\right) = -\frac{\tau}{\ln(2)}\ln\left(\frac{q'(T)}{q'(0)}\right) = -\frac{5568 \mp 30}{\ln(2)}\ln\left(\frac{13.5}{1.69}\right) = -16,692 \mp 90$$

years. The accuracy of the 14C dating process depends on a knowledge of the exact ratio of radio-

active 14C to the carbon in the atmosphere. This ratio has changed over the years. There is the basic sinusoidal variation about the mean with a period of roughly 8,000 years. In addition, volcanic eruptions, and industrial smoke produce more 14C into the atmosphere and decrease the ratio. The most drastic change has occurred because of testing of nuclear weapons resulting in an increase of 100% in the ratio in some parts of the Northern Hemisphere. These events change the ratio and therefore in living tissue. These variations have to be factored into the dating process. The same method can be used for short term dating by using material with shorter half-lives. For example, white-lead, a pigment used by painters has a half-life of only 22 years. Similarly using material such as uranium with a half-life in the billions of years we can date material of much longer periods.

Appendix G: Probability Theory Foundations

An **experiment** is a procedure that yields one of a given set of possible outcomes. The **sample space** of the experiment is the set of possible outcomes. An **event** is a subset of the sample space. Laplace's definition of the probability of an even with finitely many outcomes is: The probability of an event E, which is a subset of a finite sample space of S of equally likely outcomes is p(E)=|E|/|S| where |.| means the cardinality of the set. The probability of the complementary event E is given by $p(\bar{E}) = 1 - p(E)$. This normalizes probability to the interval [0,1] so that the probability of the occurrence of an event *or* the non-occurrence is 1 which is the probability of a "certain" event since one of them must occur. An event with probability 0 is an **impossible** event; an event with the probability 1, is a **certain/sure** event. We can then use the inclusion exclusion principle to show that the probability of a union of two events is:

F.1) $$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$$

If two events are mutually exclusive then the probability of their occurrence simultaneously is zero so that the last term is zero. If two events can occur simultaneously then we have conditional probability where the conditional probability of the conditioned event E_c given the event E_g is defined;

F.2) $$p(E_c|E_g) = \frac{p(E_c \cap E_g)}{p(E_g)}$$

From the above we can see that $p(E_c \cap E_g) = p(E_c|E_g) \cdot p(E_g)$. Therefore if the equality $p(E_c \cap E_g) = p(E_c) \cdot p(E_g)$ holds we say that the events are independent. It should be noted that this is not related to "mutual exclusivity". This property of independence cannot be shown using Venn diagrams and it is a waste of time to attempt it. Even if two events are mutually exclusive they are not independent since then we know that if one of the events occurred the probability of the occurrence of the other is zero. The definition of probability is derived from the idea of "the limiting value of the relative frequency of occurrence" of an event. The implication is that probabilities are defined for specific events from statistical considerations of mass phenomena or repetitive phenomena. In other words, in order to apply the theory of probability, we must have practically an unlimited sequence of uniform observations. We simplify the theory and the empirical process by considering only some of the essential properties (attributes) of ness phenamena oe repetitive events. Probability theory, then, deals with the construction of a rational theory based on the simplest possible exact concepts, which although admittedly is inadequate to represent the complexity of the real processes, reproduces some of their essential properties. In this sense, the probability of occurrence of a 6 or 1 of a given die is a property analogous to its mass, specific heat or electrical resistance. The theory of probability is concerned with relations existing between physical quantities of this kind. The axiomatic foundations or the theory of probability rest on a few basic postulates.

i) The set of points representing the possible outcomes of an experiment is called the sample space or the event space of the experiment; denoted by n.

ii) Each outcome in the sample space Is called a sample point and is denoted by w. Associated with each sample point w_i is a non-negative real number $p_i < 1$. The p_i can be understood as the probability of occurrence of the elementary event associated with the sample point w_i.

iii) The probability that an event E, will occur is the sum of the probabilities of the sample points that are associated with the occurrence of E. tie can write this as;

F.3a) $$P(E) = \sum p_i$$

Suppose we throw a single die. What is the probability of a 3? The sample space consists of all the die's faces of which there are six. A single side (such as 3) has a probability of 1/6. We write this as P(that a 3 faces up)=1/6, or P(3)=1/6. An alternative definition of P(E) is

F.3b) $$P(E) = \lim_{n \to \infty} \frac{m}{n}$$

where m= number of successful occurrences of the event E in n trials, n = number of times an experiment is performed. The basic assumption is that P(E), which mary be understood to be a property of a process or an entity (physical or otherwise), is the uniform limit of an unlimited sequence of operations. Before going on to a more rigorous exposition of the fundamentals of probability theory it would be beneficial to further discuss the elements of randomness upon which probability theory rests. The essential difference between a sequence of results obtained in a random event and a deterministic event consists in the possibility of devising a method of selecting the elementary elements so as to produce a fundamental change in the relative frequencies. The relative frequencies (probabilities) associated with the random events of a random process do not change If the partial sampling sequences are changed. This Impossibility of affecting the chances of a process by a system of selection is the characteristic and decisive property common to all sequences of observation and mass phenomena. Simply put, (a) Relative frequency of an attribute must have a limiting value, and (b) This value must remain the same in all partial sequences which may be selected from the original one in an arbitrary way.

The preceding discussion may be formalized by basically paraphrasing Kolmogorov [Kolmogorov,1956] who first laid the mathematical foundations of probability theory. As a result of Kolmogorov's creation, Probability Theory joined the rest of the exact mathematical disciplines. The Axioms of Probability Theory may be summarized as follows.

Let W be a collection of elements $(w_1,w_2,\cdots,w_n)$ which we call the elementary events and E the set of subsets of W; the elements of the set E are called random events.

i) E is a field of sets,
ii) E contains the set W.
iii) To each set E_i in E is assigned a non-negative real number $P(E_i)\cdot$ This number $P(E_i)$ is called the probability of the event E_i.
iv) P(W) =1
v) If E_i and E_j have no elements in common (i.e. are mutually exclusive), then

$$P(E_i + E_j) = P(E_i \cup E_j) = P(E_i) + P(E_j)$$

A system of sets E, together with a definite assignment of the numbers $P(E_i)$ satisfying Axioms (i) through (v) is called a field of probability. A simple field of probability can be constructed as follows. Take an arbitrary finite set $W= \{w_1,w_2,\cdots,w_n\}$ and an arbitrary set $\{p_1,p_2,...,p_n\}$ of non-negative numbers with the sum

F.4) $$p_1 + p_2 + \ldots + p_n = 1$$

E is the set of all subsets of W and $E_i \in E$ consists of the set of points $\{w_{i1}, w_{i2}, \cdots, w_{in}\}$ such that

F.4) $$P(\omega_{i1}, \omega_{i2}, \ldots, \omega_{i3}) = p_{i1} + p_{i2} + \ldots + p_{in}$$

In this case $p_{i1} + p_{i2} + \ldots + p_{in}$ are the probabilities of the elementary events $\omega_{i1}, \omega_{i2}, \ldots, \omega_{i3}$ or simply *elementary probabilities*. In this way, we derive all possible finite fields of probability in which E consists of the set of all subsets of W.

Addition Theorem: Logical-OR of probabilities

If two events E_1 and E_2 are combined *additively* to form a new event then

F.5) $$P(E_i + E_j) = P(E_i \cup E_j) = P(E_i) + P(E_j) - P(E_i \cap E_j)$$

$P(E_1+E_2)$ is the probability of occurrence of either event E_1 or event E_2, and corresponds to the union of the two sets E_1 and E_2. If the events are mutually exclusive (a pair of disjoint sets), then

F.6) $$P(E_i + E_j) = P(E_i \cup E_j) = P(E_i) + P(E_j)$$

As of yet nothing has been said about $P(E_1E_2)$ which is the probability of occurrence of events E_1 and E_2 together and is the intersection of the two sets E_1 and E_2. The two expressions above are sometimes known as the Addition Theorem of Probability and they are Included In the basic Axioms of Probability due to Kolmogorov,

Multiplication Theorem: Logical-AND of probabilities

The concepts of *independence of events*, *conditional probability* and the so-called Multiplication Theorem can be derived from set theory and must be understood together within the context of probability theory as they are related to each other. Consider two events E_1 and E_2 In the sample space W, where E_1 represants an event that has occurred, and E_2 represents an event whose occurrence or non-occurrence is of interest. Also $P(E_1) > 0$. Then, the conditional probability of the occurrence of event E_2, given that event E_1 has occurred, $P(E_2|E_1)$ is defined to be

F.7) $$P(E_2|E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} = \frac{P(E_1E_2)}{P(E_1)}$$

this is also known as the Multiplication Theorem of Probability. The multiplication is used as an equivalent symbolism for the intersection, the same way that addition was used for union earlier. The extension of eq. (F.7) to many events is as follows:

F.8) $$P(E_1E_2\ldots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2)\ldots P(E_n|E_1E_2\ldots E_3)$$

Then, two events E_1 and E_2 are said to be independent if

F.9) $$P(E_2E_1) = P(E_1 \cap E_2) = P(E_1)P(E_2)$$

The implication of the above equation is that

F.10) $$P(E_2|E_1) = P(E_2)$$

The concept of independence may also be extended to any number of events:

F.11) $$(E_1E_2\ldots E_n) = P(E_1)P(E_2)P(E_3)\ldots P(E_n)$$

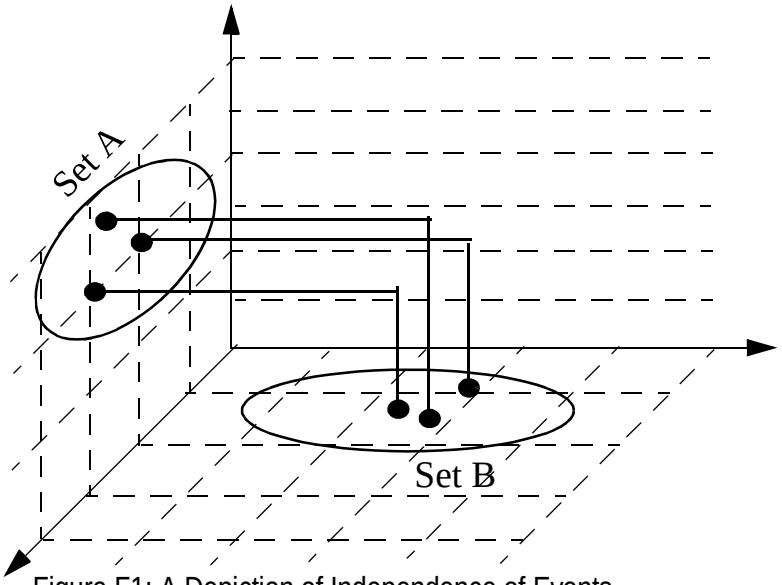


Figure F1: A Depiction of Independence of Events

The concept of independence occurs in probability theory and is a concept different than that of mutual exclusivity since mutual exclusivity is about disjointness of sets. We can define independence of sets by allowing many universes (which is also standard practice in logic since a universe of discourse is always specified). As can be seen the concept of independence of sets and different universes is a concept that can be added to sets (actually a concept that can be added to Venn diagrams) as shown above. This diagram also makes clear what it means to obtain the Cartesian product of two sets. We can represent such products by increasing dimensions of the space (or the dimensions of the Venn Diagrams). The difficulty arises because a set by itself is a single dimensional concept, and a Venn diagram is already a two dimensional construct. Therefore a Cartesian product of two sets (which is a two dimensional concept) becomes entangled with the two dimensionality of the standard Venn diagram. Much of the difficulty of attempting to display pictorially probability independence apart from mutual exclusivity is due to the denial of the single dimensionality of the set concept and its representation as a two dimensional object instead of relating it to the number line where it naturally belongs.

Suppose we throw two dice. What is the probability that the sum of the dice will be 11? To do this we need to construct the sample space. In this case it is probably easiest to construct a matrix (a 2-dimensional table) which lists all the possibilities and from these we can create another 6×6 table

which shows the sum of the dice faces.

F.13)

$$\begin{bmatrix} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{bmatrix} \qquad \begin{bmatrix} 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 6 & 7 & 8 & 9 & 10 \\ 6 & 7 & 8 & 9 & 10 & 11 \\ 7 & 8 & 9 & 10 & 11 & 12 \end{bmatrix}$$

We simply count the cells in which the integers sum to 11 and that happens to occur in only 2 places out of 36. Since the sample space consists of 36 possibilities and the favorable outcomes are 2, the probability of 11 is 2/36. We note that this is considered to be two independent trials since what happens to one die does not affect what happens to the other die. We also assume that the dice are fair in that the outcomes are equiprobable. For example, the probability that the sum of the dice rolls will be divisible by three can also be counted off the second matrix to be 12/36.

Suppose a protolanguage consists strings length 3 at random from the alphabet with one vowel and one consonant(stop) A={i,p} with repetetion.. Let E_1 be the event that the string of length 3 begins with the vowel *i*. Let E_2 be the event that there are an even number of *p's* in the string. Are E_1 and E_2 independent if the 8 words of this language are equally likely? There are $8=2^3$ strings of length three. The strings which begin with the vowel *i* are *ipp, iip, ipi, iip, iii*. The number of strings with an even number of p's in them are *iii*, and *ppi, pip, ipp*. Therefore $P(E_1)=5/8$, and $P(E_2)=4/8$. Therefore $E_1 \cap E_2 =\{\text{iii,ipp}\}$ and $P(E_1 \cap E_2)=2/8$. If the events had been independent we'd have $P(E_1 \cap E_2)=P(E_1)P(E_2)=(5/8)(4/8)=20/64$.

Appendix H: Phonological Phase Space

There is an approximate 3-D phonological/space phase space [Hubey,1994]) derived via dimensional analysis which can be used when referring to discussion of phonological/phonetic issues such as distance, resemblance, etc. It can be seen in Figure G.1. Many common phenomena of linguistics can be visually seen in this figure. Mathematical reasons as to why it works can be seen in Hubey[1994], and in more detail in Hubey[1999a]. In this figure it is clear why /ptksn/ are rarely absent in languages. The /p/ is the extreme X (except for /w/); /k/ is the practical extreme for X (minimum) and /n/ defines the minimum Z. Any smaller value in the Z direction than /n/ would fall in the vocalic group.

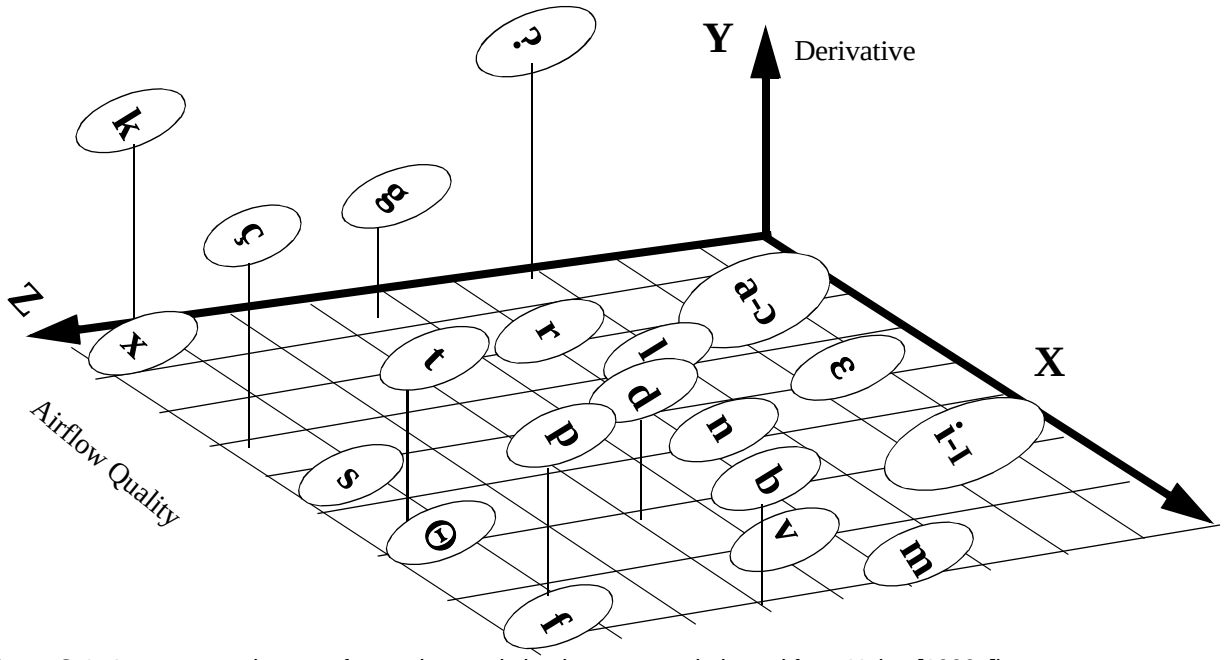


Figure G.1: A representative set of speech sounds in phase space (adapted from Hubey[1999a]).

Although the figure is not scaled, it should be noted that, in general, the relative positions of the phonemes do not change appreciably, however it's not possible without more evidence/data to be able choose a more exact placement of the sounds, however, many of the concepts of phonology can be explicated via this diagram. A concept that we'll need the discussion of lenition/fortition is that of a *vector*. An interesting usage of the concept of vectors will applied to lenition as can be seen in Lass [Lass, 1984, p. 177] that gives the phonological rules for lenition and fortition as;

(a) Stop > Fricative > Approximant > Zero
(b) Voiceless > Voiced

Essentially the same results can be found in Foley [1977]. These results can easily be shown to be derivable quite clearly and unambiguously in the phase space and are related to sonority. We only need two dimensions although three would be better) and the concept of a Cartesian vector to show the essential results. The space shown in Figure G.2 is a 2-dimensional subspace of the 3-dimen-

sional phase spaces developed earlier. Indeed, the three-dimensional phase space can be considered to be a subspace of the many different feature-bundle spaces discussed in the literature with the caveat that these spaces are not orthogonal and the mapping might not be one-one or linear.

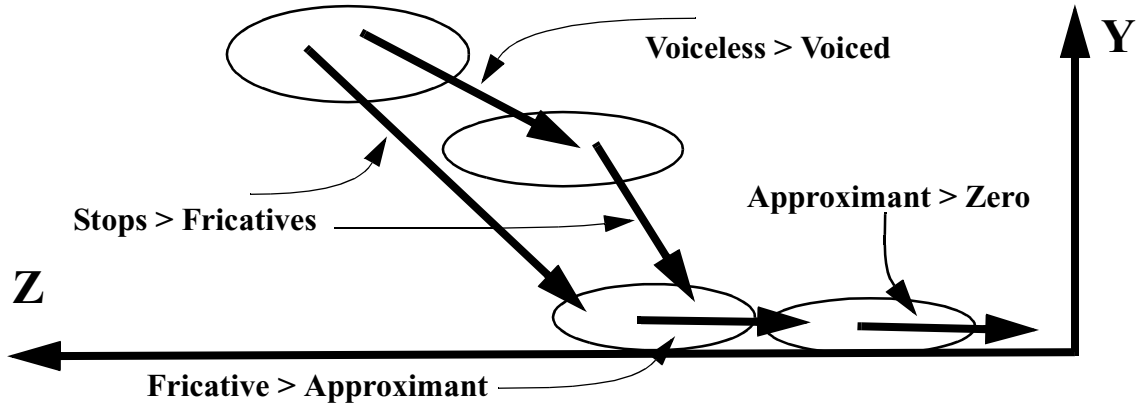


Figure G.2 Fortition, Lenition and Sonority These concepts of weakening or strengthening of sounds are very fundamental in phonology [see for example Foley,1977]. The basic ideas can easily be "explained" in terms of the vector phase space of this papers. Figure G.1 has been altered very slightly to exaggarate the effects for better and discrimination.

We can see immediately from Lass's hierarchy that in Fig. G.2 the vectors point in the negative Y direction (e.g Stop > fricative) or negative Z directions (e.g. Fricative > Approximant). In other words, the vectors that depict these relationships (lenition) point toward the origin. Since no measurements have been taken to indicate the scale of the phase space, and no mathematical definitions have been given, at best we can use the data from Lass[1984] and Foley[1977] as guides to make the phase space reflect reality as closely as possible. One can also give an excellent explanation of other phenomena such as child language development, aphasia, vowel placements, sonority in 3-dimensions, diphthongs, birth of new phonemes, etc [Hubey,1999a], Hubey,1994].

Similarly a 2-dimensional 'explanation' of sonority can be seen to pop right out of the 3-dimensional phase space. The idea of a sonority scale can be explained, first, directly from the graph since the sonority scale seems to extend from the vowels toward the plosives so the scale is essentially the distance from the origin of the axis, the voiceless plosives being the least sonorant and the low vowels being the most sonorant; thus being inversely proportional to the distance from the origin (at least in the two dimensions as shown). The sonority, a related concept, then can be expressed as a number of different ways using mathematical relationships [Hubey,1999a].

References

Anttila, R. (1989) Historical and comparative linguistics, John Benjamins Pub. Co., Amsterdam

Barnsley, M. (1988) Fractals Everywhere, Academic Press, New York.

Baxter, William, and Alexis Manaster Ramer, forthcoming Review of Ringe (1992). Diachronica (personal communication)

Berg, H.C. (1993) Random Walks in Biology, Princeton University Press, Princeton, NJ.

Bender, M. (1969] Chance CVC correspondences in unrelated languages, Language, No. 45, 1969, pp.519-531.

Benninga, S. (1998) Financial Modeling, MIT Press, Cambridge, MA.

Bryis, E., M. Bellalah, H. Mai, and F. De Varenne (1998) Options, Futures, and Exotic Derivatives, John Wiley, New York.

Campbell, L.(1998) Historical Linguistics: An Introduction, The MIT Press, Cambridge, MA.

Clark, J. and C. Yallop (1990) Phonetics and Phonology, Blackwell, Oxford.

Cowan, H. (1962) Statistical determination of linguistic relationships, Studia Linguistica 16, pp.57-96.

Crowley, T. (1997) An Introduction to Historical Linguistics, Oxford University Press, NY.

Dixon, R. (1997) The Rise and Fall of Languages, Cambridge University Press, New York.

Doerfer, G. (1971) Khalaj Materials, Indiana University Press, Bloomington, IN.

Einstein, A. (1956) Investigations on the Theory of the Brownian Movement, Dover Publications, New York.

Feller, W. (1957) An Introduction to Probability Theory and its Applications, Vol I, Wiley and Sons, New York.

Foley, J., Foundations of Theoretical Phonology, Cambridge Univ. Press, Cambridge, 1977

Fox, A. (1995) Linguistic Reconstruction: An Introduction to Theory and Method, Oxford University Press, NY.

Gardiner, C.W. (1983) Handbook of Stochastic Methods, Springer-Verlag, New York.

Graham, R., D. Knuth and O. Patashnik(1989) Concrete Mathematics, Addison-Wesley, Reading, MA.

Greenberg, J. (1960) A qualitative approach to morphological typology of language, IJAL, vol 26, No. 3, July 1960, pp.179-194.

Greenberg, J. (1993) Observations concerning Ringe's "On calculating the factor of chance in language comparison". Proceedings of the American Philosophical Society, 137(1): 79-90.

Gronbech, K. (1979) The Structure of Turkic Languages, Indiana University Press, Bloomington, IN

Gronbech, V. (1979) Preliminary Studies in Turkic Historical Phonology, Indiana University Press, Bloomington, IN.

Hauser, M. (1997) The Evolution of Communication, MIT Press, Cambridge, MA.

Hubey, H.M., K. Sivaprasad and R. Vasudevan (1983) , *Scattering from a Random Slab*, Journal of Radio Science, March-April 1983.

Hubey, H.M. (1999a) Vector Phonological Spaces via Dimensional Analysis, Journal of the International Quantitative Linguistics Association, accepted for publication.

Hubey, H.M. (1999b) *Logic, Fallacy, Reasoning and Superficial Resemblance*, submitted to History of Language

Hubey, H.M. (1998a) Quantitative Methods in Linguistics with an Application to *PIE, History of Language, September 1998.

Hubey, H.M. (1994) Mathematical and Computational Linguistics, Mir Domu Tvoemu, Moscow, Russia.

Hubey, H.M.(1998b) Intellectual Tools: minimum mathematics for the social sciences, to be published.

Hubey, H.M. (1979) Deterministic and Stochastic Behavior of Continuous Production Systems, MS Thesis, NJIT, New Jersey.

Hubey, H.M, and D. Deremer (1999) Dynamic Weighted Averages: Nonlinearity and Fuzzy Logic in Prediction, International ICSC Congress on Computational Intelligence: Methods and Applications, June 22-25, Rochester Institute of Technology, Rochester, NY.

Jazwinski, A. (1970) Stochastic Processes and Filtering Theory, Academic Press, New York.

Kolmogorov, A.N. (1956) Foundations of Probability Theory, Chelsea Publishing, New York.

Lass, R. (1984) Phonology, Cambridge University Press, Cambridge.

Lass, R. (1997) Historical linguistics and language change, Cambridge University Press, New York.

Lieberman, P, and S. Blumstein (1988) Speech Physiology, Speech Perception and Acoustic Phonetics, Cambridge University Press, Cambridge.

Manaster-Ramer, in Joseph C. Salmons and Brian D. Joseph (eds) (1998) Nostratic. Sifting the Evidence, Benjamins, Amsterdam.

Ringe, D. (1992) On Calculating the Factor of Chance in Language Comparison, The American Philosophical Society. Transactions of the American Philosophical Society, vol. 82, Philadelphia.

Ringe, D. (1995) Nostratic and the Factor of Chance, Diachronica XII:1.55-74.

Ruhlen, M. (1994) On the origin of languages : studies in linguistic taxonomy, Stanford University Press, Stanford, CA.

Soong, T.T. (1973) Random Differential Equations in Science and Engineering, Academic Press, New York.

Stevens, S.S. (ed) (1966) *Handbook of Experimental Psychology,* John Wiley & Sons, New York.

Stratonovich, R.L. (1967) Topics in the Theory of Random Noise, Gordon and Breach, New York.

Srinivasan S.K. and R. Vasudevan (1971) Introduction to Random Differential Equations and Their Applications, American Elsevier, New York.

Torgerson, W. (1958), Theory and Methods of Scaling, Wiley and Sons, New York.

Topçuoglu, et al (1996) Dictionary of the Turkic Languages, Routledge, New York.

USN&WR (1998) Baby Talk, June 15.