Beyond lumping and splitting: probabilistic issues in historical linguistics

William H. Baxter
Alexis Manaster Ramer

Symposium on Time Depth in Historical Linguistics, 19–22 August 1999
McDonald Institute for Archaeological Research, Cambridge

1. Introduction

Unlike synchronic linguists, who can ask their consultants for additional data and examples, historical linguists are, for the most part, stuck with the data we have. Only rarely are we able to add to the corpus of available texts, or find additional examples of a rare combination of phonemes. The problems of historical linguistics are thus inherently probabilistic: we must constantly decide whether perceived patterns in our data are significant or simply due to chance. Actually, this is true at any time depth; but the importance of probabilistic issues is especially clear when considering possible remote linguistic relationships.

Unfortunately, the traditions of historical linguistics equip us poorly to handle these probabilistic issues. Our discipline had already achieved impressive scientific results in the nineteenth century, before modern significance-testing procedures were developed. Lacking reliable techniques for extracting maximal information from minimal data, and having plenty of other things to do, historical linguists have tended to focus on problems where meaningful results seemed assured, and have avoided investing time in riskier endeavours. Though the inherent interest and seductive appeal of the remote linguistic past are undeniable, pushing back the frontiers of time has generally not been a high priority for those actually working in historical linguistics.

In this paper, we argue that the temporal reach of historical linguistics can probably be extended significantly in some cases, but only if the probabilistic issues are faced squarely, and new techniques of handling them are developed and widely understood. As an illustration, we propose a way of dealing with a *Gedankenexperiment* recently proposed by Hock & Joseph (1996, 491–3): if we had data only from Modern English and Modern Hindi, without evidence from other Indo-European languages, would we be able to detect the genetic relationship between them? Hock & Joseph express pessimism; but as we shall see, the problem yields to a fairly simple probabilistic technique. This example encourages us that our field will be able to penetrate deeper into the linguistic past.

2. The logic of linguistic time-depth

What time depths are attainable in historical linguistics? We can imagine at least three kinds of possible answer to this question, as in (1):

(1) Limits on time depth: three kinds of answers

1 There is a general limit on time depth which is a property of the methods of historical linguistics; this limit cannot be exceeded, regardless of the details of any particular investigation.

2 Attainable time depth will always be limited in any particular investigation, but the limit varies from one investigation to another, depending on the available information, and on the cleverness and luck of the investigators.

3 There are no limits on attainable time depth in historical linguistics.

On the first view, the limit on time depth is a general one, like the speed of light or Planck's Constant. This view seems to be widespread among historical linguists, though it is rarely argued for explicitly, and we know of no good evidence or arguments for it (except for uninteresting time depths, such as times prior to our emergence as a species). We suspect that its apparent popularity results in part from the failure to distinguish between answers 1 and 2: if one rejects answer 3 (as most sensible people would), then at first glance, answer 1 may seem the only alternative.

According to answer 2, the one we prefer, the limits on attainable time depth in linguistics are analogous to those in genealogical research on family history. Since evidence tends to deteriorate over time, information about the past is always incomplete, and there will come a point when investigation produces no further results. But the limit is not a general, *a priori* one; it is specific to the situation. The only way to learn whether one has reached this limit is to dig deeper, and evaluate the results; there is no reason, still less a duty, to stop digging, as long as there is a chance of getting lucky.

The difference between answers 1 and 2 can be seen as one of logical form—specifically, scope of quantification. Informally, if *S* is a variable over research situations in historical linguistics, and *T* is a variable over moments of time, then answer 1 can be paraphrased as in (2):

(2) (**There exists** a time *T*) such that (**for every** research situation *S*),
no investigation in *S* can give reliable results about times earlier than *T*.

Answer 2, on the other hand, has the quantifying expressions in the reverse order:

(3) (**For every** research situation *S*) (**there exists** a time *T*) such that
no investigation in *S* can give reliable results about times earlier than *T*.

Of the two formulations, (2) could justify a general argument against claims of early relationship:

(4) Claim *C* makes assertions about times earlier than *T*.
By (2), no such claim can be reliable.
Therefore, the claim *C* must be unreliable.

But notice that this argument does not go through for the formulation in (3). From (3), one can argue that *in any given situation*, there will always be time depths beyond our reach; but there is no way of knowing, a priori, where that limit is; each particular claim must be evaluated on its own merits. Only if we accept the formulation (2) does it make sense to argue about the value of the time-depth limit *T* in general; if we accept (3) instead, then what is needed is rather a way of evaluating one's results—in other words, a significance test. In the next section, we discuss why techniques for such testing have been slow to develop in historical linguistics.

3. Historical linguistics: a crisis in the immune system

The traditions and academic microcultures of historical linguistics present a number of obstacles to the development of appropriate mathematical approaches to probabilistic questions: (1) The techniques linguists call 'the comparative method', though marvellously useful, were formulated before modern significance testing had been developed, and do not adequately handle probabilistic problems. (2) Historical linguists generally receive little mathematical training, and are poorly equipped, as a group, to develop or evaluate probabilistic approaches. (3) The discourse of historical linguistics still reflects inductivist and positivist attitudes to scientific method which prevailed in the nineteenth and early twentieth century (especially in social and behavioural science). These attitudes include an aversion to hypothesis-making, and a belief that proper scientific inquiry gradually increases the store of certain knowledge by applying fixed rules of induction to observations and to previously 'proven' results.

From a sociological point of view, any academic discipline needs a way to protect itself from unbridled speculation: a way to steer its practitioners away from hypotheses which are likely to be both unproductive and contentious, and a rationale for avoiding lengthy discussions with enthusiastic but uninformed amateurs. Moreover, certain beliefs—quite apart from their truth or falsity—can be useful as badges of legitimate membership in a profession: to distinguish the astronomers from the astrologers, the physicians from the faith healers, the physicists from the metaphysicists, the scientists from the pseudo-scientists. The famous ban on discussions of language origins by the Société de Linguistique de Paris[1] was probably intended as such a protection; indeed, the whole positivist ideology of science perhaps served such a function. In the absence of better significance-testing procedures in historical linguistics, the belief in a general limit on attainable time depth may have served a useful function in the past. We suggest, however, that this traditional defence mechanism and others like it have outlived their usefulness. They are like broad-spectrum antibiotics, which attack a wide range of bacteria in order to protect against a subset of harmful ones. In our view, historical linguistics should now develop more sensitive and discriminating techniques of hypothesis testing, which can protect us from meaningless or untestable speculations on the one hand, without precluding investigation of deep linguistic relationships on the other.

The conditions for developing these techniques are not ideal, however. Although graduate students in some areas of linguistics (such as phonetics and sociolinguistics) are expected to have at least a basic cookbook-style course in statistics, this is less common in historical linguistics. Traditionally, historical linguists are expected to learn several difficult ancient languages as well as a sprinkling of modern ones, and this honourable tradition is to be treasured; but it leaves little time for even that exposure to science and mathematics which is routinely expected of social scientists. (In the US, at least, the problem is compounded by the heavy pressure put on graduate programs to get PhD students through the pipeline in five years, regardless of their field of study.)

Those historical linguists who do not shrink from the mathematical questions that arise in their work usually rely on a combination of self-instruction and help from more knowledgeable colleagues. But even when conscientiously applied, these strategies often fail to protect against what in other fields would (we assume) be considered elementary mistakes. The protections of the peer-review process often fail, as well.

It is not hard to illustrate the difficulty we historical linguists have, as a group, with mathematical arguments. Don Ringe is to be commended for perceiving the importance of probability and tackling it head-on. But it remains true that his 1992 study shows, not just minor technical

mistakes, but (as we believe) a basic misunderstanding of hypothesis testing (see Baxter & Manaster Ramer 1996; Ringe 1998, 186–7 and *passim*; Baxter 1998). The same is true of Ringe (1995), which attempts to debunk the Nostratic theory by fitting the distribution of Illič-Svityč's Nostratic etymologies to a binomial distribution (never mind which one). This argument was chosen in a widely-used undergraduate probability course called CHANCE[2] as an example of how *not* to do statistics (Snell 1995; Baxter 1998, 218n). Yet both works continue to be cited approvingly for their 'elegant' and 'devastating' arguments (Lass 1997, 169n), or for their 'mathematical proofs' (Campbell 1997, *passim*).

As pointed out by Manaster Ramer & Hitchcock (1996), the problem is by no means limited to believers in a general limit on time-depth: in a popular article in *Scientific American*, Greenberg & Ruhlen (1992, 98) fall victim to the old birthday-problem fallacy (for further discussion see also Snell, Paterson & Grinstead 1998). Nor are we ourselves immune: when already preparing camera-ready copy for *A Handbook of Old Chinese Phonology* (1992), Baxter discovered a serious flaw in a crucial probabilistic argument, which took several weeks to repair. (He has confidence in the published version, but in any case the whole section—pages 97–137—was largely ignored by reviewers.)

It might be thought that a poor understanding of probability is a problem only when investigating distant linguistic relationships. But in fact, it is not great time depth, but scarcity of evidence, which necessitates probabilistic reasoning; and scarcity of evidence is a problem historical linguists face at all time depths. In deciding whether there are enough examples of a phonological correspondence to justify a sound law, or sufficient evidence from daughter languages to reconstruct a particular form in their common ancestor, one uses probabilistic reasoning, good or bad. Deciding such matters by traditional rules of thumb rather than by mathematical argument is a relic of the nineteenth-century origins of our discipline.

Finally, the idea that our methods allow us to 'prove' language relationships to a certain limit, beyond which the responsible scientist must refrain from speculation, reflects a nineteenth-century inductivist ideology of science which is now rightly discredited. In the inductivist view, scientists carefully observe facts, their minds uncontaminated by preconceived notions or hypotheses; and they prove new scientific results by applying a fixed code of valid inductive principles to their observations. (A 'method' in the narrow sense, as in 'the comparative method', is a code of this kind.) As long as scientists unswervingly follow this procedure, it is believed, the truth of their results is assured, and the store of legitimately proven scientific knowledge is gradually increased. But speculations not firmly grounded in observation undermine the legitimacy of the whole process, and pollute the inquiry from that point on.

This inductivist view, though too rigid to follow in practice, and now largely abandoned by philosophers of science, still survives among the defence mechanisms of our field. By suggesting that hypotheses about deep linguistic relationships are forever beyond the reach of legitimate scientific inquiry, it is now doing a disservice by unnecessarily and prematurely discrediting some of the most interesting lines of inquiry open to us. We urgently need more discriminating defences which will protect us without exacting this high price.

4. Hindi and English: a probabilistic approach illustrated

A problem suggested by Hock & Joseph in their recent textbook of historical linguistics (1996) will help to show the power of probabilistic methods in extracting meaningful patterns from very incomplete data. Hock & Joseph consider what would happen if we tried to evaluate

the relationship between Modern English and Modern Hindi, without the benefit of the other Indo-European evidence which makes it clear that they are genetically related. Having pointed out that 'clearly related languages can come to be different enough that many of their genuine cognates are difficult or even impossible to recognize', they go on:

> Let us pursue this issue a little further by taking a closer look at the relationship between Modern Hindi and English – pretending that we do not yet know that they are related, and trying to establish their relationship by vocabulary comparison. This is actually more difficult than it appears. It is all too easy to be influenced by one's knowledge of the historical relationship between the two languages and therefore to notice the genuine cognates, or even to underestimate the effects of linguistic change on the recognizability of genuine cognates. (1996, 491)

The approach they choose to test is to search Hindi and English dictionaries for possible cognates. Without other evidence, it is of course difficult to identify genuine cognates this way. Many cognates have been so affected by sound change that their relationship is not apparent (e.g. 'horn', English [hɔrn], Hindi [sĩːg]), while there are many chance lookalikes which we know, from other evidence, to be unrelated (e.g. 'cut', English [kʌt], Hindi [kaːʈ-naː]). They conclude that matches found by such a dictionary search would be true cognates only about half the time. These 'dismal' results (as they say) lead them to doubt that a more distant relationship, such as that proposed between Indo-European and Uralic, could ever be convincingly established (1996, 493).

Hock & Joseph's pessimism results in part from their lack of clear criteria for 'dismalness'. Note that they do not actually try to estimate the probability that the resemblances they find between English and Hindi could be due to chance. On the question of how much evidence would be required to build a convincing case, they comment:

> Clearly, one correspondence is not enough; nor are twenty. And just as clearly, a thousand correspondences with systematic recurrences of phonetic similarities and differences would be fairly persuasive. Are 500 enough, then? And if not, are 501 sufficient? *Nobody can give a satisfactory answer to these questions.* And this is no doubt the reason that linguists may disagree over whether a particular proposed genetic relationship is sufficiently supported or not. (1996, 493; emphasis added)

This is, in fact, the major point of our paper: as long as historical linguists 'can[not] give a satisfactory answer' to questions about the significance of their data, there will be little consensus on which distant relationships are real and which illusory. On the other hand, if we can develop reliable answers to such questions, then the temporal reach of historical linguistics can probably be extended.

We would like to offer a different approach to the English-Hindi problem, to illustrate how probabilistic procedures can work. The procedure uses controlled word lists and explicit criteria for identifying phonetic matches; the number of matches obtained when words are paired by meaning is compared with the result when they are systematically paired at random. The null hypothesis to be tested is that the score obtained when pairing words by meaning is not significantly greater than the scores obtained when pairing words at random.[3]

The basic elements of the procedure are (1) a predetermined list of basic word meanings to be tested, and (2) an explicit algorithm for deciding when to count a pairing as a phonetic

match. Words are chosen from two languages to match the chosen set of basic word meanings. In the present case we use a 33-word list, given in (5) below with English and Hindi pronunciations:[4]

(5) 33-word list

	meaning	English	Hindi
1	blood	[blʌd]	[kʰuːn]
2	bone	[boʊn]	[haɖɖiː]
3	to die	[daɪ]	[mar-naː]
4	dog	[dɔg]	[kuttaː]
5	ear	[ɪr]	[kaːn]
6	egg	[ɛg]	[anɖaː]
7	eye	[aɪ]	[ãːkʰ]
8	fire	[faɪər]	[aːg]
9	fish	[fɪʃ]	[matʃʰliː]
10	full	[fʊl]	[bʰaraː]
11	to give	[gɪv]	[de-naː]
12	hand	[hænd]	[haːtʰ]
13	horn	[hɔrn]	[sĩːg]
14	I	[aɪ]	[mãĩ]
15	know	[noʊ]	[dʒaːn-naː]
16	louse	[laʊs]	[dʒũː]
17	moon	[mun]	[tʃãːd]
18	name	[neɪm]	[naːm]
19	new	[nju]	[najaː]
20	one	[wʌn]	[eːk]
21	salt	[sɔlt]	[namak]
22	stone	[stoʊn]	[pattʰar]
23	sun	[sʌn]	[suːradʒ]
24	tail	[teɪəl]	[pũːtʃʰ]
25	this	[ðɪs]	[je]
26	thou	[ju]	[tuː]
27	tongue	[tʌŋ]	[dʒiːbʰ]
28	tooth	[tuθ]	[dãːt]
29	two	[tu]	[doː]
30	water	[ˈwatər]	[paːniː]
31	what	[hwʌt]	[kjaː]
32	wind	[wɪnd]	[havaː]
33	year	[jɪr]	[saːl]

This list is adapted from a list of 35 especially basic word-meanings (given in Starostin 1991, 59–60) chosen by S.E. Jaxontov from the 100-word Swadesh list.[5] From Jaxontov's list, we omit the meaning 'nose' because of the possibility that phonetic symbolism might favour nasal consonants. Also, since 'who' and 'what' are similar in many languages (including English), they should probably not be considered independent choices; we include only 'what'. We chose this list because it is basic (thus relatively resistant to borrowing) and conveniently short; any similar list would do, as long as it is not biased in some way for or against the null hypothesis.

The procedure is to count the number of phonetic matches between the English and Hindi lists: first when the list items are paired by meaning, and then (many times) when they are paired at random. We will use these results to estimate the probability that random pairing could produce as many matches as we obtained when pairing the words by meaning.[6] The procedure fits the general pattern of a significance test in inferential statistics: we test the null hypothesis that the matches obtained when pairing words by meaning are no more than would be expected when pairing words at random. We do so by measuring how extreme the observed number of matches is, given the predictions of the null hypothesis. And we set our level of significance—our criteria for 'extremeness'—in advance: we will reject the null hypothesis if the probability *P* of getting the observed number of matches (or more) is less than .05, or less than one in 20.

When identifying phonetic matches, it is important that we use explicit criteria which can be mechanically applied (e.g., by a computer). For the present test, we use a very general algorithm involving similarity of initial consonants only: words match if their initial consonants belong to the same one of the 10 classes of consonants defined by Dolgopolsky (1964; 1986), listed in (6) below:

(6) The 10 'Dolgopolsky classes'

	type	description
1	P	labial obstruents [p, b, f]
2	T	dental obstruents (except hissing and hushing sibilants)
3	S	[s, ʃ, z, ʒ]
4	K	velar and postvelar obstruents [k, g, x] and affricates such as [ts, tʃ, dz, dʒ]
5	M	[m]
6	N	[n], [ɲ], and noninitial [ŋ]
7	R	[r, l]
8	W	[w] and initial [u]
9	J	[j]
10	Ø	laryngeals, zero consonant, and initial [ŋ]

Dolgopolsky's intention was to define the classes so that consonants in the same class are more likely to change, over time, to another member of the same class than to some consonant of another class; for us, this classification is simply a way to define a very rough (but explicitly defined) relationship of phonetic similarity which may have historical relevance. For convenient identification, each class is labelled with a capital letter, as in (6) above.

Each of the 33 words in English and Hindi is assigned to one of the 10 classes as in (7) below; matches—where English and Hindi words are in the same class—are in boldface type.

(7) English and Hindi 33-word lists, with Dolgopolsky classes (matches in bold)

	meaning	English	class	Hindi	class
1	blood	[blʌd]	P	[kʰuːn]	K
2	bone	[boʊn]	P	[haɖḍiː]	Ø
3	to die	[daɪ]	T	[mar-naː]	M
4	dog	[dɔg]	T	[kuttaː]	K
5	ear	[ɪr]	Ø	[kaːn]	K
6	**egg**	[ɛg]	**Ø**	[anɖaː]	**Ø**
7	**eye**	[aɪ]	**Ø**	[ãːkʰ]	**Ø**
8	fire	[faɪər]	P	[aːg]	Ø
9	fish	[fɪʃ]	P	[matʃʰliː]	M
10	**full**	[fʊl]	**P**	[bʰaraː]	**P**
11	to give	[gɪv]	K	[de-naː]	T
12	**hand**	[hænd]	**Ø**	[haːtʰ]	**Ø**
13	horn	[hɔrn]	Ø	[sĩːg]	S
14	I	[aɪ]	Ø	[mãĩ]	M
15	know	[noʊ]	N	[dʒaːn-naː]	K
16	louse	[laʊs]	R	[dʒũː]	K
17	moon	[mun]	M	[tʃãːd]	K
18	**name**	[neɪm]	**N**	[naːm]	**N**
19	**new**	[nju]	**N**	[najaː]	**N**
20	one	[wʌn]	W	[eːk]	Ø
21	salt	[sɔlt]	S	[namak]	N
22	stone	[stoʊn]	S	[pattʰar]	P
23	**sun**	[sʌn]	**S**	[suːradʒ]	**S**
24	tail	[teɪəl]	T	[pũːtʃʰ]	P
25	this	[ðɪs]	T	[je]	J
26	thou	[ju]	J	[tuː]	T
27	tongue	[tʌŋ]	T	[dʒiːbʰ]	K
28	**tooth**	[tuθ]	**T**	[dãːt]	**T**
29	**two**	[tu]	**T**	[doː]	**T**
30	water	[ˈwatər]	W	[paːniː]	P
31	what	[hwʌt]	Ø	[kjaː]	K
32	wind	[wɪnd]	W	[havaː]	Ø
33	year	[jɪr]	J	[saːl]	S

Thus, when these English and Hindi words are paired by meaning, our algorithm defines the following nine pairs as phonetic matches, since in each case their initial consonants belong to the same Dolgopolsky class:

(8) English-Hindi matches from the 33-word lists

	meaning	English	class	Hindi	class
6	egg	[ɛg]	Ø	[anɖaː]	Ø
7	eye	[aɪ]	Ø	[ãːkʰ]	Ø
10	full	[fʊl]	P	[bʰaraː]	P
12	hand	[hænd]	Ø	[haːtʰ]	Ø
18	name	[neɪm]	N	[naːm]	N
19	new	[nju]	N	[najaː]	N
23	sun	[sʌn]	S	[suːradʒ]	S
28	tooth	[tuθ]	T	[dãːt]	T
29	two	[tu]	T	[doː]	T

This matching algorithm is rather crude, of course, and like Hock & Joseph's dictionary search, it is not very good at identifying actual cognates. Of the nine matches above, only six—'egg', 'name', 'new', 'sun', 'tooth', and 'two'—are actually connected etymologically.[7] Also, the algorithm is too crude to identify such true cognates on the list as 'horn' (English [hɔrn], Hindi [sĩːg], both from an Indo-European root *k^{j}er-), and 'know' (English [noʊ], Hindi [dʒaːn-naː], both from Indo-European *$g^{j}enh_3$-).

The next step in our procedure is to estimate the likelihood that a score this high (nine or more matches out of a possible 33) could have occurred by chance. In order to do this, we need to know the probability of each of the possible outcomes when the list items are paired at random.[8] It is difficult to compute these probabilities directly, but it is easy (and intuitively straightforward) to estimate them by using a computer simulation: we program a computer to run repeated trials in which the words of the English and Hindi lists are paired at random (rather than by meaning); and for each such random pairing, we count the number of phonetic matches between paired words, using the same algorithm as when they were paired by meaning. If we perform many such trials and keep track of the score for each trial, we can see how often we get nine (or more) matches. The frequency with which this happens is an estimate of the probability that nine (or more) matches would occur by chance.

We have performed such a simulation, using the scripting language HyperTalk, part of the HyperCard software package for Apple Macintosh.[9] One thousand trials were run, as described above. For concreteness, the results of a typical trial are given in (9) below. (Only the letters representing Dolgopolsky classes are listed, not the full forms.) Notice that the 'English class' column has the same letters, in the same order, as in (7) above; the 'Hindi class' column has the same letters as in (7) above, but in scrambled order. If the *i*-th letter in the English column is the same as the *i*-th letter in the Hindi column, that is counted as a match. In this particular trial, there were four matches: at items 8, 14, 19, and 22.

(9) A typical random English-Hindi pairing (matches in bold)

	meaning	English class	Hindi class	match?
1	blood	P	K	false
2	bone	P	T	false
3	to die	T	Ø	false
4	dog	T	M	false
5	ear	Ø	K	false
6	egg	Ø	P	false
7	eye	Ø	N	false
8	**fire**	**P**	**P**	**true**
9	fish	P	S	false
10	full	P	T	false
11	to give	K	M	false
12	hand	Ø	M	false
13	horn	Ø	K	false
14	**I**	**Ø**	**Ø**	**true**
15	know	N	Ø	false
16	louse	R	T	false
17	moon	M	K	false
18	name	N	K	false
19	**new**	**N**	**N**	**true**
20	one	W	Ø	false
21	salt	S	Ø	false
22	**stone**	**S**	**S**	**true**
23	sun	S	J	false
24	tail	T	K	false
25	this	T	Ø	false
26	thou	J	N	false
27	tongue	T	P	false
28	tooth	T	S	false
29	two	T	K	false
30	water	W	T	false
31	what	Ø	K	false
32	wind	W	P	false
33	year	J	Ø	false

One thousand trials of this kind were performed; the results are summarized in (10):

(10) Results of 1000 trials, pairing the English and Hindi lists at random

15	trials with	0	matches	
59	trials with	1	match	
124	trials with	2	matches	
197	trials with	3	matches	
255	trials with	4	matches	
187	trials with	5	matches	
89	trials with	6	matches	above the line:
42	trials with	7	matches	trials with scoresof **less than 9 matches**
21	trials with	8	matches	(989 trials of 1000)
8	trials with	9	matches	below the line:
3	trials with	10	matches	trials with scores of **9 or more matches**
0	trials with	11	matches	(11 trials of 1000)
0	trials with	12	matches	
1000	trials in all			

The 1000 random trials produced scores from 0 to 10 matches; the most common single score was four matches (255 trials). Recall that we got nine matches when the words were paired according to meaning; this score was equalled or exceeded in 11 of the 1000 trials (eight trials with nine matches, three trials with 10 matches). We can use this result to estimate the probability that nine or more matches would occur by chance in this situation: the probability is about 11/1000, or .011. This is well below our previously chosen significance level of $P < .05$, indicating that we should reject the null hypothesis (that the observed matches between English and Hindi could be the result of chance).[10]

These results agree well with the analysis of such experiments given in Justeson & Stephens (1980). The expected probabilities are difficult to calculate directly, but it can be shown that they are closely approximated by the function given in (11). This function is an example of a *Poisson density function*, used in probability theory to predict the outcome of a variety of processes, including the number of radioactive atoms that will decay within a certain time interval, or the number of telephone calls that will come into an exchange in a certain length of time (Hoel, Port & Stone 1971, 56–7). In this case, P[*k*] is the probability that a particular random trial will produce a score of *k* matches; *n* is the number of words on the list (33 in our case); *r* is the total number of possible pairings which the match-recognition algorithm will identify as matches (in this case, 128);[11] and *e* is the base of the natural logarithms:

$$P[k] = e^{-r/n} \frac{(r/n)^k}{k!}$$

(11)

The Poisson density is usually described in terms of a single parameter 'λ' (lambda), as in (12) below; in our case, lambda is the ratio *r/n*, or 128/33. (We can think of this ratio as a measure of

the looseness of the match-recognition algorithm: the greater the value of lambda, the higher the probability that words will match when paired at random.)

$$P[k] = e^{-\lambda} \frac{\lambda^{k}}{k!} \quad (12)$$

The actual values of this function, with lambda = 128/33, are given in (13).

(13) Probability of *K* matches in a single trial (Poisson approximation, $\lambda = 128/33$):

K	P [*k* = *K*]	P [*k* ≤ *K*]	
0	0.020676	0.020676	
1	0.080197	0.100873	
2	0.155534	0.256407	
3	0.201095	0.457502	
4	0.195001	0.652503	
5	0.151273	0.803777	
6	0.097793	0.901569	above the line:
7	0.054188	0.955758	outcomes where $k < 9$:
8	0.026273	0.982031	P [$k < 9$] = .982
9	0.011323	0.993354	below the line:
10	0.004392	0.997746	outcomes where $k \geq 9$:
11	0.001549	0.999295	P [$k \geq 9$] = 1 – .982 = .018
12	0.000501	0.999795	
...	...	...	

In the graph of Fig. 1, the results from our 1000 random trials are displayed and compared with the outcomes predicted by the Poisson approximation. The vertical bars indicate the actual frequency of each score in the 1000 random trials (as in (10) above); the black squares (connected by a line) indicate the expected values according to the Poisson approximation (as in (13) above). The 'observed result' of nine matches is the number of matches obtained when the lists were paired by meaning (as in (7)); a vertical line through the graph separates the outcomes which are as 'extreme' as this observed value (the right side of the graph, sometimes called the 'tail') from those that are less extreme (the left side). Of the 1000 trials, 11 are to the right of this line, and 989 to the left; so (as in (10)) our simulation indicates that the probability of getting a score of nine or more in a random trial is about 0.011. By the Poisson approximation (as in (13)), the probability of getting a score less than nine is 0.982031, so the probability of getting a score of nine or higher is 1 – 0.982031, or about .018. Both estimates (.011 and .018) fall well below the previously chosen significance level of .05, indicating that we should reject the null hypothesis: that is, we should conclude that the observed score is not the result of chance.[12]

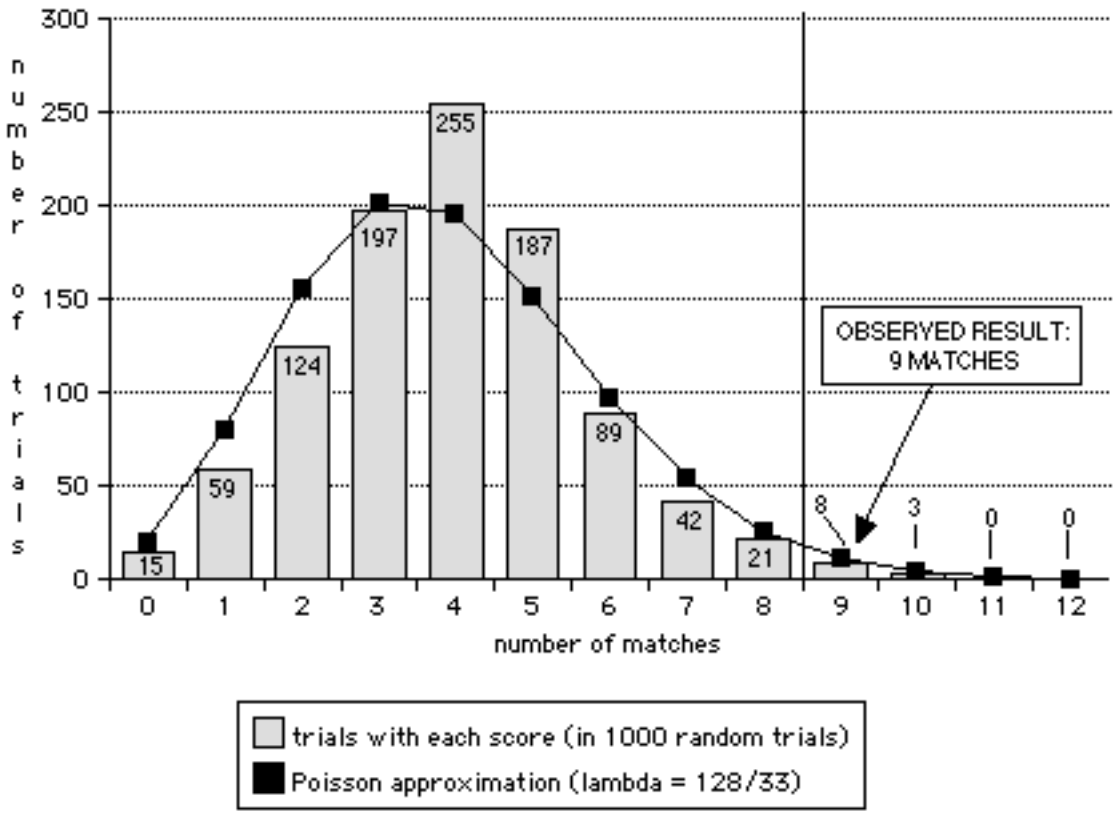

Figure 1. *Results of English-Hindi comparison: pairing by meaning compared with 1000 random pairings*

Let us pause to consider the significance of this result. Our test involved only 33 test items from each language, and we used only very general information about the initial consonants of those items. We have not worked out the actual phonological correspondences between English and Hindi, much less reconstructed their common ancestor (though some valid correspondences begin to emerge even in this small sample, including English [t] ↔ Hindi [d], [n] ↔ [n], [m] ↔ [m]). Yet on this evidence alone, we appear to be in a position to offer odds of, say, fifty to one that the two languages are genetically related. How did this come about?

It is probably the few relatively well-preserved cognate forms like ‘name’, ‘new’, ‘tooth’, and ‘two’ in our sample which produce this result. While it is true that sound changes sometimes make cognates unrecognizable, and chance can create deceptive lookalikes, it is not uniformly so, and in this case enough scraps are left to give a statistically significant result when the data are handled with due caution. We suspect that by using careful statistical inference to squeeze maximal information out of our data, it may well be possible to identify remote linguistic relationships which are not now widely accepted. Even in cases where the data may be too sketchy for detailed etymological analysis, what we learn, when combined with the findings of other disciplines, may help answer important questions about prehistory.

5. The simulation procedure: final comments

We close with a few further comments on the simulation procedure illustrated here.

1. *On the looseness or strictness of the match-recognition algorithm* . A strict algorithm is one that, all other things being equal, produces a small number of matches (a small value of *r*, as detailed in note 9); a loose algorithm gives a greater number of matches. The algorithm used here is fairly loose: it identified nine matches, only some of which are true cognates, between the English and Hindi lists when paired by meaning. It might seem that such looseness is necessarily a bad thing: will we not be fooled into thinking that the evidence for a relationship is stronger than it actually is? But our procedure protects us from this danger. We draw no conclusions from raw scores in isolation, but only from comparing them with the range of scores obtained in repeated random trials using the same algorithm. If the algorithm is loose, the scores will be higher in both cases; if the algorithm is strict, the scores will be lower. The principle is the same as with standardized scores for tests like the Graduate Record Examination: the questions in one version of the test may be harder or easier than in another version, so raw scores are not comparable; instead, we use standardized scores which evaluate each result in terms of its place in the total distribution of scores on that version of the test.

Also, just as a test can be either too easy or too difficult to give useful information, in our procedure, a matching algorithm can be either too strict or too loose to be useful. The loosest possible algorithm would allow anything to match anything; in that case, every possible pairing (whether by meaning or not) would score 33 matches out of a possible 33. On the other hand, suppose we used an algorithm which allowed an item on the first list to be matched only by a total reduplication of itself on the second (e.g. 'tongue' could be matched only by 'tonguetongue'). With this algorithm, not only English and Hindi, but even English and itself would always produce a score of zero. To be useful, an algorithm must fall somewhere between these extremes.

2. *On circularity*. Any algorithm for identifying matches can be used in our procedure, as long as it can be mechanically applied and always gives an answer, yes or no. (This is, in fact, the definition of an algorithm: an explicit procedure which always stops and gives an answer.) The algorithm used here is based on broad similarity of initial consonants, but there is nothing essential in this; more specific algorithms could be devised, which look at every segment, look for certain particular correspondences, and so forth. Likewise, our algorithm examines initial consonants only; but other algorithms could be designed which are more (or less) sensitive to phonotactic positions.

There is, however, a potential danger of circularity: one way to get an apparently significant result is to design the algorithm to fit the particular word lists being used. For example, the English-Hindi word lists happen to show two cases ('dog' and 'tongue') where English words of the T-class are paired with Hindi words of the K-class. Building this T-K correspondence into the match-recognition algorithm would boost the score obtained when words are paired by meaning, and might (mistakenly) be taken as evidence in favour of a relationship.[13] One way of testing more specific hypotheses, while at least minimizing circularity, is to incorporate observed correspondences into the algorithm, as long as they are justified by sufficiently many examples outside the list being tested.[14]

3. *On failing the test*. What if the test had 'failed'—that is, what if the observed score had not been high enough to discredit the null hypothesis? It is a point often misunderstood, and worth emphasizing, that *a negative result is not evidence against a genetic relationship*; nor is it evidence that no relationship is 'demonstrable'. A negative result means only that one particular

way of identifying matches, applied to one particular set of items, gives no positive evidence for a relationship. It is no indication of whether the data themselves are 'random', for data are not random or nonrandom in themselves, but only with respect to some (explicit) probabilistic model of a random process.[15]

The procedure illustrated in this paper is just one way of handling one kind of problem; the general technique can be extended to a variety of situations, not just the detection of distant relationships. As noted earlier, the problem of judging sufficiency of evidence arises constantly in historical linguistics, at all time depths.[16] Besides being intuitively clear, computer simulation can be a powerful technique for investigating the predictions of null hypotheses which may be too complex to be handled by traditional formulae. If we can develop and perfect probabilistic techniques appropriate to the problems of historical linguistics, we may perhaps have success finding answers which once seemed beyond our reach.

NOTES

[1]'ARTICLE PREMIER. — La Société de Linguistique a pour but l'étude des langues, celle des légendes, traditions, coutumes, documents, pouvant éclairer la science ethnographique. Tout autre objet d'études est rigoureusement interdit.
'ART. 2. — La Société n'admet aucune communication concernant, soit l'origine du langage, soit la création d'une langue universelle.' (Société de Linguistique de Paris 1871, iii)

[2]The course's web page may be found at `http://www.dartmouth.edu/~chance/`.

[3]This technique was illustrated in Baxter (1995) for reconstructed Tibeto-Burman and Old Chinese; it is essentially the same as the procedure analyzed by Justeson & Stephens (1980). Oswalt's 'shift test' (1998) is a technique of the same general kind; see Baxter (1998, 222–8) for discussion of his application of it to Indo-European and Altaic.

[4]Different choices are possible in some cases, but we know of no alternatives which would substantially affect the outcome of the test. Peter Hook has pointed out to us that most Hindi speakers would probably find [zabaːn] a more natural choice than [dʒiːbʰ] for 'tooth'. We accordingly performed a new 1000-trial simulation, with [zabaːn] substituted for [dʒiːbʰ], and arrived at essentially the same result; see notes 10 and 12 below for details.

[5]Jaxontov divided the 100-word list into 35 more basic and 65 less basic words. His idea was that genetically related languages would normally have a greater percentage of cognates among the 35 more basic words than among the 65 less basic ones, while the reverse pattern might indicate that shared items were due to borrowing.

[6]For consistency, we use 'pairing' and 'match' as follows: The match-recognition algorithm takes an English-Hindi word-pair as input, and returns 'true' if the pair are a phonological match, 'false' otherwise. Each trial involves applying the algorithm to 33 word-pairs; the score for the trial is the number of pairs (out of the 33) for which the algorithm returns a value of 'true'. We call each such pair a 'match'; we use 'pairing' to refer to the process of choosing which 33 word-pairs will be tested in a given trial.

[7]Moreover, of these six items, Hindi [suːradʒ] 'sun' is a loan from Sanskrit (as a questioner at the Symposium pointed out), and thus would not count as an English-Hindi cognate.

[8]One of the serious problems with the procedure described in Ringe (1992) is that he provides no way of calculating this distribution, and thus no way of knowing whether a particular observed outcome is significant or not (see Baxter & Manaster Ramer 1996, 376–7, 383). His attempts to clarify the issue in a more recent paper (1998, 157–61) are confused and unsuccessful; he still provides no coherent model of the random process being modelled, and fails to show that his test statistic would be relevant, even if its distribution were known (Baxter 1998, 228–30).

[9]The HyperTalk script for running a single random trial is given below by way of illustration. Though the programming language may be unfamiliar to many readers, its conventions are fairly intuitive, and the comments (preceded by a double hyphen '--') should make it possible to follow the general process.

```
function oneTrial
  --
  -- This routine runs a single trial in which
  -- the words of two lists, A and B, are paired at random
  -- and the number of phonetic 'matches' are counted.
```

```
    -- The details of each trial are recorded on a newly created card in
    -- the stack.
    --
    -- Note:  "bg" = "background" (of a card)
    --        "fld" = "field" (of a card), for storing numbers or text
    --        "it" is a variable filled by the "get" command
    --
    go bg "trials" -- (the appropriate background for the new card)
    doMenu "New Card" -- create a new card
    put "" into bg fld "results" -- initialize the "results" field
    get bg fld "totalTrials" -- keep track of which trial this is
    put it into bg fld "trialNumber"
    --
    -- Read the two word lists from data cards elsewhere in the stack
    --
    put AList() into Alines -- read list 'A' from data cards
    put BList() into Blines -- read list 'B' from data cards
    put the number of lines of Alines into numOfItems
    --
    -- Make a random list of numbers 1 to 'numOfItems'.
    -- This allows us to scramble the order of items on the second list:
    -- if "r(i)" is the i-th item on the random list, we pair the
    -- i-th item of list A with the r(i)-th item of B.
    --
    put makeRandomList(numOfItems) into randomList
    put "" into resultHolder -- a temporary variable to keep the results
    put 0 into numOfMatches -- initialize the variable that counts matches
    repeat with i = 1 to numOfItems -- for each item on list A
      put i into message box -- (show the user which item is being done)
      put line i of Alines into Aforms -- get the i-th item of list A
      get item i of randomList -- get "r(i)", the i-th random number
      put line it of Blines into Bforms -- get the r(i)-th item of list B
      get ABmatchLists(Aforms, Bforms) -- true if they match, else false
      if it is true then add 1 to numOfMatches -- increment counter of matches
      put numOfMatches into bg fld "numOfMatches" -- show new value on card
      get Aforms && "|" && Bforms && it & return -- make a record of this pair
      put it after bg fld "results" -- add to cumulative record for this trial
      put it after resultHolder
    end repeat
    --
    -- update the card where the results of trials are accumulated
    --
    go cd "tally"
    put numOfMatches + 1 into lineNum -- 1st line is 0 matches
    get item 2 of line lineNum of cd fld "results"
    add 1 to it
    put it into item 2 of line lineNum of cd fld "results"
    add 1 to cd fld "totalTrials"
  end oneTrial
```

[10]In the 1000-trial simulation using Hindi [zabaːn] instead of [dʒiːbʰ] for 'tooth', there were likewise 8 trials with 9 matches and 3 trials with 10 matches, for a total of 11 of 1000 random trials in which the number of matches was 9 or more.

[11]To be more precise: Let E be the set of words on the English list, and H the set of words on the Hindi list; each set has 33 elements. The *Cartesian product* of the two sets, $E \times H$, is defined as the set of all ordered pairs $<e,\ h>$, where e is a word from the English list (*not* the base of the natural logarithms, this time), and h a word from the Hindi list; this set $E \times H$ contains 33 x 33 or 1089 such pairs as elements. In the formula in (11), the number r is the number of ordered pairs $<e,\ h>$ in $E \times H$ where e and h match each other according to the match-recognition algorithm. For the lists and algorithm used here, $r = 128$. The reader can verify this by making a table of 33 rows and columns, where each row is labelled with a letter from the English classes, and each column with a letter from the Hindi classes. Let the cell at the intersection of the i-th row and the j-th column be one, if the i-th class-letter on

the English list is the same as the *j*-th class-letter on the Hindi list, and zero otherwise. Of the 1089 cells in the table, 128 will contain a one.

[12]When we use Hindi [zaba:n] (Dolgopolsky class 'S') instead of [dʒi:b^h] (Dolgopolsky class 'K') for 'tooth', the value of lambda in the Poisson approximation becomes 130/30 rather than 128/33, yielding a probability of about .020 of getting nine or more matches by chance—still well below the significance level of .05.

[13]Something similar happened with Ringe's comparison of English and Turkish (1992, 47–51), which seemed to produce a false positive result because, in the 100-word Swadesh lists, there happened to be six cases of English /b/ corresponding to Turkish /k/. Presumably this correspondence is a chance characteristic of this particular sample of words, and would not be borne out in larger or differently chosen samples; in any case, as Ringe shows, the words cannot actually be cognates. But since his procedure looks for correspondences within the 100-word sample, and then attempts to evaluate those correspondences using the very same sample, circularity is built into his procedure, and the effects of chance can be misinterpreted.

[14]This approach is used in Baxter (1995, 4–24).

[15]The argument in Ringe (1995) involves a misunderstanding of this point.

[16]For example, Baxter (1992) used probabilistic techniques to test whether patterns found in Old Chinese rhyming were significant enough to require an explanation other than chance.

REFERENCES

Baxter, W.H., 1992. *A Handbook of Old Chinese Phonology*. Berlin: Mouton de Gruyter.

Baxter, W.H., 1995. 'A stronger affinity ... than could have been produced by accident': a probabilistic comparison of Old Chinese and Tibeto-Burman, in *The Ancestry of the Chinese Language*, ed. W.S.-Y. Wang. (Journal of Chinese Linguistics Monographs, 8). Berkeley (CA): Project on Linguistic Analysis, 1–39.

Baxter, W.H., 1998. Response to Oswalt and Ringe, in Salmons & Joseph 1992, 217–236. [A response to Oswalt 1998 and Ringe 1998.]

Baxter, W.H. & A. Manaster Ramer, 1996. Review of Ringe 1992. *Diachronica* 13(2), 371–384.

Campbell, L., 1997. *American Indian Languages: the Historical Linguistics of Native America*. New York: Oxford University Press.

Dolgopolsky, A.B., 1964. Gipoteza drevnejšego rodstva jazykovyx semej severnoj Evrazii s verojatnostnoj točki zrenija. *Voprosy Jazykoznanija* 1964(2), 53–63.

Dolgopolsky, A.B., 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia, in *Typology, Relationship, and Time: a Collection of Papers on Language Change and Relationship by Soviet Linguists*, eds. V.V. Shevoroshkin & T.L. Markey. Ann Arbor (MI): Karoma, 27–50. [A translation of Dolgopolsky 1964.]

Greenberg, J.H. & M. Ruhlen, 1992. Linguistic origins of Native Americans. *Scientific American* 267 (November), 94–99.

Hock, H.H. & B.D. Joseph, 1996. *Language History, Language Change, and Language Relationship: an Introduction to Historical and Comparative Linguistics*. Berlin: Mouton de Gruyter.

Hoel, P.G., S.C. Port & C.J. Stone, 1971. *Introduction to Probability Theory*. Boston (MA): Houghton Mifflin.

Johnson, G., 1995. Linguists debating deepest roots of language. *The New York Times* 27 June 1995, C1.

Justeson, J.S. & L.D. Stephens, 1980. Chance cognation: a probabilistic model and decision procedure for historical inference, in *Papers from the 4th International Conference on*

Historical Linguistics, eds. E.C. Traugott, R. Labrum & S. Shepherd. (Current issues in Linguistic Theory, 14). Amsterdam: Benjamins, 37–46.

Lass, R., 1997. *Historical Linguistics and Language Change*. Cambridge: Cambridge University Press.

Manaster Ramer, A. & C. Hitchcock, 1996. Glass houses: Greenberg, Ringe, and the mathematics of comparative linguistics. *Anthropological Linguistics* 38(4), 601–620.

Oswalt, R.L. 1998. A probabilistic evaluation of North Eurasiatic Nostratic, in Salmons & Joseph 1998, 199–216.

Ringe, D.A., Jr., 1992. *On Calculating the Factor of Chance in Language Comparison*. Philadelphia: The American Philosophical Society.

Ringe, D.A., Jr., 1995. 'Nostratic' and the factor of chance. *Diachronica* 12, 55–74.

Ringe, D.A., Jr., 1998. Probabilistic evidence for Indo-Uralic, in Salmons & Joseph 1998, 153–197.

Salmons, J.C. & B.D. Joseph (eds), 1998. *Nostratic: Sifting the Evidence*. (Amsterdam Studies in the Theory and History of Linguistic Science, Series IV: Current Issues in Linguistic Theory, 142). Amsterdam: John Benjamins.

Snell, J.L., 1995. Linguists debating deepest roots of language. *CHANCE News* 4(10). Available `http://www.dartmouth.edu/~chance/chance_news/recent_news/chance_news_4.10.html#Linguists`, 14 June 1999.

Snell, J.L., B. Paterson & C. Grinstead, 1998. Coincidences and linguistics. *CHANCE News* 7(6). Available `http://www.dartmouth.edu/~chance/chance_news/recent_news/chance_news_7.06.html#Coincidences and linguistics`, 14 June 1999.

Société de Linguistique de Paris, 1871. Statuts approuvés par décision ministérielle du 8 mars 1866. *Bulletin de la Société de Linguistique de Paris* 1(1), iii–iv.

Starostin, S.A., 1991. *Altajskaja Problema i Proisxoždenie Japonskogo Jazyka*. Moscow: «Nauka».