JOHN HEWSON

RECONSTRUCTING PREHISTORIC LANGUAGES ON THE COMPUTER: THE TRIUMPH OF THE ELECTRONIC NEOGRAMMARIAN

1. THE PROJECT

The prime principle lying behind all comparative linguistics is the regularity of sound change. Without this principle, comparative linguistics would be mere empty speculation: any ad hoc rule could be created on the spur of the moment to justify the most fanciful etymology. But with the principle of the regularity of sound change comparative linguistics becomes a rigorous science:[1] it is possible to propose a hypothesis, and then demonstrate clearly from the data whether the hypothesis works or not. In classic scientific method if the hypothesis does not work, it is to be abandoned; if it does work it must be shown to apply to all of the data in a completely coherent fashion. All apparent anomalies must therefore be either explained as the effects of some other cause: as a result of analogy, or borrowing or dialectal interference.

The so-called neogrammarians who developed and elaborated this principle of the regularity of sound change a hundred years ago were guilty, like most revolutionaries, of pushing their doctrine to an extreme position. Today we are aware that sound change is a very complex phenomenon and not at all the somewhat simplistic regular shifting from proto language to daughter language that they imagined. Every language is, for a start, a collection of different dialects, and the same must be considered true of those proto languages we try to describe in our reconstructions. We must, as well, rid ourselves of the simplistic notion of a proto language being a centre from which, at some select and special moment of history, all the daughter languages began

[1] It is possible for a science to be rigorous without being *exact*; comparative linguistics is not an exact science. An exact science deals with mechanical laws, a rigorous science deals with observed regularities.

to diverge like the spokes of a wheel. Our knowledge of phyla and families shows us that the splitting off is much more haphazard, and that there can be contact and interference long after the original dialectal split.

Nevertheless, the principle of the regularity of sound change remains of fundamental importance, and will remain of fundamental importance in the discipline of comparative linguistics. And it follows that if sound change is regular then it should be possible to make use of machine methods in the reconstruction of proto languages. It was with such a plan in view that we set to work in the summer of 1971 to devise computer strategies that would enable us, from an input of four daughter languages, to reconstruct the lexicon of a proto language and produce a proto language dictionary on the computer from the raw input of the daughter languages.

We chose four fairly closely related Amerindian languages to work with: Fox, Cree, Menomini and Ojibwa, from which we hoped to reconstruct a lexicon of Proto Algonkian. The correspondences of these four languages had been originally worked out some fifty years ago by LEONARD BLOOMFIELD in an article in the first volume of *Language*, now famous for its solitary footnote on the universality of the principle of sound change. Bloomfield later refined this work in his masterly sketch entitled *Algonquian* in the volume *Linguistic Structures of Native America* edited by HOIJER. In short, the sound system of Proto Algonkian had been long before thoroughly worked out, so that we were in a position to enter all the reflexes and correspondences into the computer programming as necessity might arise. But, and this was the factor of interest, although the sound system had been worked out for the proto language, no one had ever applied Bloomfield's correspondences and reflexes to the totality of the data of the four daughter languages from which he worked, so that only a few hundred words of Proto Algonkian had ever been reconstructed, in spite of the lapse of over half a century. Reconstruction is, of course, enormously time consuming, and a dictionary of a proto language normally takes many years to produce.

Our aim, therefore, was to devise a programme that would use the correspondences and reflexes of the daughter languages to find cognates in an input of raw data in these languages and to reconstruct proto forms for these cognates. The proto forms and the cognates from which they had been generated would then be assembled on the computer in the form of a proto language lexicon or dictionary.

2. THE APPROACH

The phonology of reconstructed Proto Algonkian has three main aspects: the consonants of which there are 12, the consonant clusters of which there are 32, and the vowels of which there are 8. Since the correspondences of the consonants and the consonant clusters in Algonkian (as in other language families) show greater regularity and simplicity than the vowel correspondences, the decision was made from the beginning to concentrate the search for cognates on the correspondences of consonants and consonant clusters alone, and to ignore the vowels. This is necessary for two reasons: a) to spread the net wider, and b) to simplify programming and reduce machine time.

The team that embarked upon this project was composed of linguists and computer personnel. The linguists envisaged a programme that would start with a word in one language and by generating the possibilities inherent in the correspondences search for a cognate in the same grammatical and semantic area in a second language. It is quite easy for such generated possibilities to run to computations of 30 or 40 forms, so that one word in Fox, for example, could generate a search for 30 or more possibilities in Cree, perhaps find one, perhaps find more than one, or perhaps draw a blank. The search would then continue on to Menomini, taking into account the possibly considerable information already accumulated from Cree in order to compute the possibilities to be sought for in Menomini, where again one might find a possible cognate, more than one, or zero. To proceed on to Oiibwa would then require all the previous information to be digested before predicting and searching for the possibilities in Ojibwa.

One of the worst aspects of this strategy (apart from its complexity) is the impossibility of preventing the machine from deciding (on the consonant correspondences, semantic and grammatical categories alone) that a word is cognate, when the linguist can tell at a glance (from the English gloss, or from the vowel correspondences) that it is not. The machine then proceeds to assimilate this false information and compounds the error by using it in its further predictions.

For example, for Fox *poohkešamwa* " he cuts it open ", the Cree consonant possibilities are only

p hk s m
p sk s m

but such a programme would find both Cree *paaskisam* " he shoots it " and *pooskosam* " he cuts it open ". On the basis of this data the programme would then generate the following possibilities for Menomini:

p hk s m
p hk hs m
p hk ²s m
p čk s m
p čk hs m
p čk ²s m

No cognates would be found in Menomini, however, and the programme would proceed to generate the following set of possibilities for Ojibwa:

p kk š n
p kk šš n
p kk nš n
p šk š n
p šk šš n
p šk nš n

For this set the following would be declared cognate: Ojibwa *paškošaan* " he cuts it down ", *pakkweešaan* " he slices off a part ", neither of which is truly cognate with either the Fox or Cree items. The end result of such a programme, therefore, would be almost nil.

When the problems with this approach became obvious, a totally different strategy was proposed by the computer personnel. In this strategy the first procedure is to extract the consonant framework of every input word in the raw data file (*R/F*) and to create a new file (*N/F*) listing these consonant frameworks. A sample of *R/F* is given in APPENDIX A, and a sample of *N/F* is given in APPENDIX B.

Secondly every possible proto form is generated from the known reflexes for the consonant framework of each word. This is done very simply, in a single pass, and the results stored in file *W/F*. On our first run, with 3,403 items in *R/F* and *N/F*, we generated 74,049 proto form potentialities (proto-projections) showing an average of over 21 possible proto-projections for every item of raw input.

Thirdly a massive alphabetical sort is carried out to collate all the proto-projections in file *W/F* in alphabetical order within identical

grammatical and semantic sets. The result of this sort is that all identical proto-projections in the same semantic and grammatical categories are collated together: in this way they may be distinguished from those proto-projections that occur only once.

Fourthly, all sets of identical proto-projections are sorted into a new file *Cx/F*, and the results of this file are printed out. A sample of this printout is given in APPENDIX C. It is remarkable that out of a total of 74,049 proto-projections only 1,305 items were transferred to *Cx/F*, these being the only items that showed identical proto-projections in another language.

A major edit is made at this point in the system by the linguist, who surveys the sets of proto-projections, and, from his background knowledge decides on the correct reconstruction and adds the vowels. This amended reconstruction is then keypunched, along with the code numbers which give access to the cognates from which it is generated. The final dictionary file (*D/F*) is then established by a programme which first enters the reconstruction, then enters the cognates by identifying them, removing them to the dictionary file and deleting them from the raw data file. A sample of the final dictionary file, which may be used as input to an automated typesetting machine, is given in APPENDIX D.

The machine used was an IBM 370/155 and the following were the statistics of the first run of data with general programming in FORTRAN but using COBOL for the sorts:

			File Size	*CPU*
(1)	*R/F*	*N/F*	3403	0. 5.52
(2)	*N/F*	*W/F*	74049	8.23.81
(3)	*W/F*	*P/F*	74049	1.51.44
(4)	*P/F*	*Cx/F*	1305	1.01.14
			Total	11.21.92

3. CONCLUSIONS

It should be emphasized that this method does not use the correspondences (or sound relationships between daughter languages) in order to predict possible cognates, but the reflexes (or sound relationships between each daughter language and the proto language). All the

possible proto forms (consonants and consonant clusters only) are generated rather than all the potential cognates. What is important is that the machine does not gather and sort information and make decisions in order to proceed to the next step: it simply generates all the possibilities in one fell swoop, and then collates identical possibilities together. This technique streamlines the whole process and, ultimately, leaves to the linguist the decision as to whether a word is cognate or not.

It is also of importance that the major edit in this system comes at a strategically important moment: after possible cognates have been collated, and all possible proto-projections listed, but before any decisions have been taken as to what is actually cognate. The machine, in short, carries out the time consuming task of seeking out the possibilities hidden in the masses of data and presenting them to the linguist. All that the linguist then has to do is to decide on the items that are genuinely cognate.

The same basic strategy may be applied to any closely related family of languages, although the programming must be closely tailored to the details of the languages themselves. The languages used should be conservative enough to retain sufficient discriminatory features and should also generally correlate with each other in syllable structure if the programming and machine time are to be kept within reasonable bounds. If one were to attempt the reconstruction of Proto-Romance, for example, it would be advisable to use Sardinian, Spanish, Italian and Rumanian, and leave the French evidence out. The loss of intervocalic consonants, of final vowels and final consonants has much reduced the materiality of cognates in French, and thereby reduced their discriminatory power. A comparison of Italian *agosto* (with its 6 phonemes) and French *août* (with its one phoneme) immediately demonstrates how much historical information is conveyed by the Italian word and how little the French word tells us.

The French data, and that of other divergent languages can be added later, in fact, in much simpler fashion once the basic proto dictionary has been established. From a given Proto-Romance form it should be possible to predict reasonably accurately what the French word (if it exists) is likely to be. If one can define an item in such a way it would be a simple matter to initiate a search for it.

The proto forms must be found, in short, by working back upstream in time. This is the difficult part of the task and is best done from the evidence of conservative languages. Once this difficult upstream work

has been done, it is possible by much easier downstream methods to correlate the evidence from the divergent languages.

In both cases there will be a residue once the obvious cases have been dealt with by machine methods. This residue can then be printed out so that the linguist can direct his efforts to the challenging problems that remain when the purely mechanical and regular items have been cleared from the data.

Such a strategy utilizes the best possible interaction and integration of man and machine. Man is poorly designed for collating masses of data, but is able to bring a wide range of knowledge and interpretation to a single item of data at any one time. The machine, on the other hand, betrays a certain inanity when it comes to matters of interpretation, since it has no knowledge of the world, and is thereby obliged to form its interpretations on mere surface appearances. The capacity of the machine to collate and search, on the other hand, may well arouse our admiration. As for the combination of man and machine, we are only just beginning to realize the potentialities that this combination has for the world of scholarship, and especially for the discipline of Linguistics.

Appendix A.

Language	*Grammatical category*	*Indian word*	*Semantic category*	*Basic gloss*		*English gloss*
KAO	II	AKKAJIKAA	PLC	SHADE	W	BE SHADY (IT IS SHADY)
KAO	II	AKKAJIKAATTEE	PLC	SHADE	W	IT'S SHADY
KAO	II	AKKAJIKAATTEESS	PLC	SHADE	W	-IN BE IN A SHADY PLACE
KAO	AI	AKKAJIKAATTEECC	PLC	SHADE	W	-IN BE IN A SHADY PLACE
KAO	NI	AKKAJIKAATTEEWI	PLP	UMBRELLA	W	-N UMBRELLA
KAO	II	AKKAJIKISSIN	PLC	SHADE	W	BE SHADED
KAO	AI	AKKAJIKCCIN	PLC	SHADE	W	BE SHADED
KAO	NI	AKKAKKICEE	PNH	COALS	W	COALS
KAO	PT	AKKAKKICEEPWEEN	OPF	COALS	W	ROAST IN COALS
KAO	TA	AKKAMAW	UBX	WATCH	W	LIE IN WAIT (AMBUSH) FOR S.O.

Appendix B.

Language	*Semantic category*	*Grammatical category*	*Consonant structure*	*Indian word*	*English gloss*	
C	MVM	AI	P P M SK	PAPAAMISKAAW	HE CANOES ABOUT (*ALONG*)	0
C	MVM	AI	P P M T J M	PAPAAMITAAJIMOOW	HE CRAWLS ABOUT (*ALONG*)	0
C	MVM	AI	P P M T P S	PAPAAMITAAPAASOOW	HE DRIVES ABOUT (*ALONG*)	0
C	MVM	AI	P P M HT	PAPAAMOHTEEW	HE WALKS ABOUT (*ALONG*)	0
C	OPC	AI	P P M W T	PAPAAMOOWATEEW	HE CARRIES A LOAD ABOUT (*ALONG*)	0
C	MVM	AI	P P M Y M	PAPAAMEEYIMOOW	GOES ABOUT THINKING WELL OF SELF (?)	0
C	OPU	AI	P P T K P Y H	PAPEETIKOPAYIHOOW	HE *DOUBLE*S HIMSELF UP, CROUCHES (*DOUBLE*)	0
C	UBX	AI	P P T K S N	PAPEETIKOSIN	HE LIES ROLLED UP (*DOUBLE*)	0
C	UBS	AI	P P T K P	PAPEETIKWAPIW	HE SITS HUNCHED	0
C	AMG	AI	P P T S	PAPEETISIW	HE IS SLOW (*COME*)	0

Appendix C.

Identity No.	*Language*	*Semantic category*	*Grammatical category*	*Consonant structure*	*(proto - pro:ections)*	*Indian word*	*English gloss*	
241	2	CMVMAI	P M J	M		PIMIJIMEEW	*HE CANOES (*SIDEWISE*)*	OSIDEWISE
1824	2	MMVMAI	P M J	M		PEMEEJEM3W	HE PADDLES OR SWIMS *ALONG*, ON, PAST	OALONG
2856	2	CMVMAI	P M K	J	N	PIMAKOOJIN	TOPPLE OVER	OTOPPLE
238	2	CMVMAI	P M K	J	N	PIMAKOJIN	*HE FLIES PAST (*ALONG*)*	OALONG
1842	2	MMVMAI	P M M	K		PEMOOMEKOW	HE RICES ON, *ALONG* ON HORSEBACK	OALONG
2868	2	CMVMAI	P M M	K		PIMOOMIKOW	RIDE HORSE*RIDE*BACK/PIGGY*RIDE*BACK	ORIDE
254	2	CMVMAI	P M ND			PIMOHTEEW	HE WALKS (*ALONG*)	OALONG
1841	2	MMVMAI	P M NC			PEMOOHN3W	HE WALKS ON, *ALONG*, PAST	OALONG
1094	3	FMVMAI	P M NSW			PEMOSEEWA	HE WALKS ALONG (*BY*)	OBY
1092	3	FMVMAI	P M NSW			PEMISEEWA	HE FLIES PAST (*BY*)	OBY

Appendix D.

O PATTAAKKOCKAAN	KNOCK, BUMP AGAINST S.T.
M P3QTAAHKOSKAM	HE BUMPS INTO IT AS A TREE, SOLID *BY ACCIDENT*
*PEQTAAXKWIHCINWA	AI BUMP (*-AAXKW)
M P3QTAAHKIHSEN	HE BUMPS INTO A TREE OR SOLID *BY ACCIDENT*
O PATTAAKKOCCIN	BUMP/KNOCK AGAINST S.T.
*PEQTAHOSOWA	AI ACCIDENTAL (*-AH)
M P3QTAHOSOW	HE CHOPS HIMSELF *BY ACCIDENT*
C PISTAHOSOOW	HE SHOOTS OR KNOCKS HIMSELF *BY MISTAKE*
*PEQTECWESOWA	AI ACCIDENTAL (*-ECW)
C PISTISOSOOW	HE CUTS HIMSELF BY ACCIDENT (*BY MISTAKE*)