AN ALGORITHM FOR IDENTIFYING COGNATES BETWEEN RELATED LANGUAGES

Jacques B.M. Guy

Linguistics Department (RSPacS)
Australian National University
GPO Box 4, Canberra 2601 AUSTRALIA

ABSTRACT

The algorithm takes as only input a list of words, preferably but not necessarily in phonemic transcription, in any two putatively related languages, and sorts it into decreasing order of probable cognation. The processing of a 250-item bilingual list takes about five seconds of CPU time on a DEC KL1091, and requires 56 pages of core memory. The algorithm is given no information whatsoever about the phonemic transcription used, and even though cognate identification is carried out on the basis of a context-free one-for-one matching of individual characters, its cognation decisions are bettered by a trained linguist using more information only in cases of wordlists sharing less than 40% cognates and involving complex, multiple sound correspondences.

I FUNDAMENTAL PROCEDURES

A. Identifying Sound Correspondences

Consider the following wordlist from two hypothetical Austronesian-like languages:

	Titia	Sese
"eye"	mata	nas
"sea"	tasi	sah
"father"	tama	san
"mother"	mama	nan
"tongue"	mimi	nen
"shellfish"	sisi	hehe
"bad"	sati	has
"to stand"	ti	se
"to come"	ma	na
"with"	mi	ne
"not"	sa	ha

Take the first word pair, mata/nas. We have no information about the phonetic values of their constituent characters, we do not know whether the same system of transcription was used in both wordlists: for all we know "a" might denotes a high back rounded vowel in Titia and a uvular trill in Sese. The only assumption allowed is that in each word list the same characters represent, more or less, the same sounds. Under this assumption, the possibility that any one character of a member of a word pair may correspond to any character of the other member cannot be discarded. Thus in the pair mata/nas Titia "m" may correspond to Sese "n", "a", or "s", and so may Titia "a", "t", "a", and "s".

We summarize the evidence for these possible correspondences in an TxS matrix, where T is the number of different characters found in the Titia wordlist, S that in the Sese wordlist. Thus the evidence afforded by the first pair, mata/nas:

	a	e	h	n	s	Sums of rows
a	2	0	0	2	2	6
i	0	0	0	0	0	0
m	1	0	0	1	1	3
s	0	0	0	0	0	0
t	1	0	0	1	1	3
Sums of columns	4	0	0	4	4	12

And by all 11 pairs:

	a	e	h	n	s	Sums of rows
a	10	0	3	9	6	28
i	2	6	6	5	5	22
m	5	3	0	12	2	22
s	3	2	7	0	2	14
t	4	1	2	2	5	14
Sums of columns	24	12	18	28	18	100

Matrix A (observed frequencies)

If character correspondences between the Titia and Sese word pairs were random the expected frequency e[i,j] of recorded possible correspondences between the ith character of the Titia alphabet and the jth of the Sese alphabet would be:

$$e[i,j] = \frac{\text{sum of ith row x sum of jth column}}{\text{sum of cells}}$$

giving a matrix of expected frequencies of possible sound correspondences:

	a	e	h	n	s	Sums of rows
a	6.72	3.36	5.04	7.84	5.04	28
i	5.28	2.64	3.96	6.16	3.96	22
m	5.28	2.64	3.96	6.16	3.96	22
s	3.36	1.68	2.52	3.92	2.52	14
t	3.36	1.68	2.52	3.92	2.52	14
Sums of columns	24	12	18	28	18	100

Matrix B (expected frequencies)

Note how the six character correspondences with the greatest differences between observed and expected frequencies give the simple substitution code used for generating Sese words from pseudo-Austronesian Titia:

Titia	Sese	Observed - Expected
m	n	5.84
s	h	4.48
i	e	3.36
a	a	3.28
t	s	2.48

B. Identifying Null Correspondences

Call the difference between the observed and the expected frequency of a character correspondence its weight (a much less primitive definition of weight is used in the actual implementation).

Take the first word pair (mata/nas) and enter into a 4x3 matrix W the weights of its 12 possible character correspondences:

	n	a	s
m	5.84	-0.28	-1.96
a	1.16	3.28	0.96
t	-1.92	0.64	2.48
a	1.16	3.28	0.96

Matrix W (weights)

Call potential of a character correspondence the sum of its weight and of the highest potential of all possible character correspondences to its right, i.e.

$$Pot(i,j) = W[i,j] + \max(Pot(i+1..m,j+1..n))$$

giving the matrix of potentials P for word pair mata/nas:

	n	a	s
m	11.60	2.28	-1.96
a	4.44	5.76	0.96
t	1.36	1.60	2.48
a	1.16	3.28	0.96

Matrix P (potentials)

The character correspondence with the highest potential is here m/n (P[1,1]=11.6). Of its possible successors, that with the highest potential is a/a (P[2,2]=5.76), itself followed by t/s (P[3,3]=2.48), which has no possible successor. Thus we have:

Titia	Sese	Potential
m	n	11.60
a	a	5.76
t	s	2.48
a	zero	

The same procedure applied to the rest of the wordlist gives the proper matches, Titia finals in polysyllabic words having been deleted when deriving the corresponding Sese words.

C. A Relative Measure of Cognation

Call index of cognation the maximum potential of a word pair divided by its number of correspondences, including null correspondences. Thus in the fictitious case of Titia and Sese the index of cognation of the pair mata/nas is 2.9 (its maximum potential, 11.60, divided by the number of correspondences, 4). Word pairs with high cognation indices are found to be more often genetically related than pairs with low cognation indices.

II CURRENT IMPLEMENTATION

A. Weights.

The difference between observed and expected frequencies does not provide a satisfactory measurement of the weight of a possible character correspondence. Several alternative measurements were tested, out of which standardized scores were retained: the weight of a character correspondence was redefined as the probability of the discrepancy between its observed and expected frequencies of occurrence not being due to chance, expressed as a z score. Where absolute frequencies of 20 and less are involved the exact probability is calculated and translated into a z score using a polynomial approximation (Abramowitz and Stegun 1970).

B. Vowel/Consonant Correspondences

Disallowing correspondences between vowels and consonants vastly improved the performance of the algorithm. No human intervention is needed to identify vowels from consonants, an improved version of an algorithm described in Suhotin 1962 being used to identify characters which represent vowel sounds. Whether consonants should be allowed to correspond to vowels is left as an option in the current implementation.

C. Iterations

Performance is again improved when word pairs showing individual character matches as computed from matrices of potentials (section IB above) are reprocessed. The weights of possible character correspondences are recomputed. This time, however, only characters in the same positions in the two words are scored as possible correspondences. Thus for instance, the first pass of the algorithm having matched the "m" of "mata" to the "n" of "nas", Titia "m" is scored in the second pass as corresponding possibly only to Sese "n". Sequences of alternate null correspondences are collapsed so as not to preclude the identification of correspondences which might have been missed in the first pass, e.g. a pair mat/mot matched in the first pass as

m	m
zero	o
a	zero
t	t

is reinput in the second pass as

m	m
a	o
t	t

Weights of possible character correspondences having thus been recomputed, a new matrix of potentials and a new cognation index is computed for each word pair. Further iterations were found to yield negligible improvements to the results obtained.

D. Improved Weights and Cognation Indices

Frequent character correspondences often yield very high z scores (up to 18.2). The presence of even one such high score in a word pair often invalidates the character-matching procedure. A number of alternative alterations to the definition of weight were tried, out of which the simplest proved best: weights beyond an arbitrary value are set to that value. Practice showed a maximum value of 3.0 to 4.0 to give the best results. This is not surprising, since there is no significant difference in the degrees of certainty corresponding to z scores of 4 and beyond.

The last improvement in the performance of the algorithm to date was brought by a redefinition of the cognation index. Once the individual character matches of a word pair have been identified from its matrix of potentials their weights are adjusted as follows:
1) Positive weights less than 1.28 (corresponding to a 90% significance level) are set to zero; negative weights and weights greater than 1.28 are left unchanged.
2) Positive weights of character-to-zero matches are set to zero; negative weights are left unchanged.

The cognation index is then defined as the sum of the adjusted weights divided by the number of matches, e.g. (an actual example from two languages of Vanuatu):

		Weight	
		Original	Adjusted
x	zero	-0.64	-0.64
a	a	3.98	3.98
b	D	1.06	0.00
a	zero	2.12	0.00
t	D	3.12	3.12
i	i	2.86	2.86
a	zero	2.12	0.00

			9.32

Cognation index: 9.32/8 = 1.165

III PERFORMANCE OF THE ALGORITHM

The algorithm as described has been implemented in Simula 67 on a DEC KL1091 and applied to a corpus of some 300 words in 75 languages and dialects of Vanuatu. Results are excellent for languages sharing 40% or more cognates, even when sound correspondences are complex. They deteriorate rapidly when lesser proportions of cognates and complex sound correspondences are involved, but remain excellent when mainly one-to-one correspondences are present. Thus for instance Sakao and Tolomako (Espiritu Santo, Vanuatu) were given as sharing 38.91% cognates (cut-off cognation index: 1.28), as against a human estimate of 41% backed by a full knowledge of their diachronic phonologies and comparisons with other related languages. Out of the 50 word pairs with the highest cognation indices only two (the 38th and the 45th) were definitely not cognate and one (the 36th) doubtful. Yet, Sakao has undergone extremely complex phonological changes, viz.:

	Tolomako	Sakao
"eye"	nata	mða
"throat"	tsalo	rlɔ
"banana"	βetali	iðɛl
"to blow"	suβi	hy
"nine"	linaratati	lønɛrɛpɛð

IV FURTHER IMPROVEMENTS

The identification of environment-conditioned phonological correspondences is the next, most obvious stage in further improving the algorithm. This problem has of course been, and is being, investigated. Difficulties arise from the fact that frequencies of possible correspondences in any given environment become too low to be handled by statistical tests. Other approaches -- inspired from chess-playing programs -- have been tried, but have proved too expensive in computer time so far. A further, much desirable, improvement is the identification of rules of metathesis. The solution to this problem appears to be subordinated to that of the discovery of context-sensitive rules.

V PURPOSE OF THE ALGORITHM

A bilingual wordlist is conceptually equivalent to a bilingual text: words of a list to sentences of a text, phonemes of a word to morphemes of a sentence, cognate pairs to segments of the same meaning, non-cognates to segments of different meanings, and the algorithm described is the present state of an attempted solution to the much more general following problem: given two texts of approximately equal lengths in two different languages, determine whether one is the translation of the other -- or both translations of a text in a third language -- wholly or in parts, and if so, establish the rules for translating one into the other.

VI REFERENCES

Abramowitz, Milton and Irene A. Stegun. Handbook of Mathematical Functions. National Bureau of Standards, 1970.

Suhotin, B.V. Eksperimental'noe vydelenie klassov bukv s pomoshchju elektronnoj vychislitel'noj mashiny. Problemy strukturnoj lingvistiki. Moscow 1962.