

Distinctive Feature Matching as a Basis for Finding Cognates

T. R. Hendrie

University of Victoria

The purpose of this paper is to describe a preliminary project to search for cognates in dictionary lists using a computer. Of equal interest is the use of a systematic examination to test the hypothesis that diachronic sound shift proceeds by the very gradual process of changing only one or two distinctive phonological features at a time. By defining the phonemes of each language under consideration in terms of distinctive features which are proper in the context of the individual inventories, it is possible to compare the segments of possible cognate pairs and measure the similarity of the forms. Once some level of phonetic similarity is established, it is then worthwhile to examine the meanings of the forms for semantic correspondence. The computer program makes possible the rapid selection of only those pairs of words which are considered close enough phonetically to warrant further inspection.

Phonological comparison of languages which are the result of historical drift from a common proto language typically reveals differences between corresponding phonemes of only one or two distinctive feature values. To illustrate the validity of the premise to this project, a cursory look at the Germanic Consonant Shift will serve as an example. Table I shows the Proto-Indo-European voiceless stops *p, *t, *k together with their respective Germanic reflexes f, θ, h, in a matrix showing some of the distinctive features that might be used to differentiate them. Members of each pair made up of the ancestral consonant and its reflex differ by a single feature value, that is, seven out of eight feature specifications are the same, while pairs not supposed to reflect

systematic sound shift, for example /p/, /t/, share six or less feature values.

	CONS	SON	ANT	COR	CONT	NAS	LAB	VOICE
p	+	-	+	-	-	-	+	-
f	+	-	+	-	+	-	+	-
t	+	-	+	+	-	-	-	-
θ	+	-	+	+	+	-	-	-
k	+	-	-	-	-	-	-	-
h	+	-	-	-	+	-	-	-

Grimm's Law: P.I.E. p t k Germanic f θ h

Table I

To take an Australian example (O'Grady n.d.:4) there is an /l/¹-
/ɭ/ correspondence between Wadjuk (WJK) and Nyungar (NYU) where,
measured in shared distinctive features, the sounds are very close.

WJK	'KY-LI'	'BILO'
NYU	<i>kerɭ</i>	<i>pirɭ</i>
	'boomerang'	'creek'

Of course, none of this is very surprising, but it does demonstrate that sounds change little by little and, more importantly for the argument at hand, similarity between phonemes which are reflexes of a proto-form can be measured in a useful way. On the other hand, there is no assurance that all sounds in a phonological system change equally gradually: in a given time period one set of phones may change by a single feature while another set may change by two or more features. Consider the hypothetical cognate

¹ This /l/ stems from the 'L' used in the analysis of Wadjuk by Moore in 1884. Because such 'pre-scientific' analyses cannot always be counted on to show all the phonemic contrasts of a language, one must be careful when ascribing phonemic and phonetic values to the symbols found in such works.

pair **paka* in Australian language A and **waka* in Australian language B which manifests the *p* - *w* correspondence attested between some Australian languages. Using the same features as those of Table I, a comparison of the second consonant of each word would show all feature values to be shared, whereas a comparison of the initials would reveal that they share only three out of eight feature values. Consequently, if the program is set so that only words with consonants differing by a single feature or less are singled out, significant correlations will be overlooked. Conversely, setting the number of shared features too low will produce such a mass of 'possible' shared cognates that anything significant will be obscured.

The 'value' level of the computer program is the parameter defining the number of shared features chosen as a minimum for consideration as cognates. This number is very important to the usefulness of the paired forms in the printout. The level has to be set low enough so that nothing significant is missed and high enough to avoid so many paired forms that no advantage is gained over manually comparing the dictionaries. The previous example from Nyangamarda and Gupapuyngu shows how the most appropriate 'level' can vary even within a single pair of cognates.

The choice of distinctive features to be used in the program obviously depends on the languages being compared and varies with the number of phonemes in each language and the similarity of their inventories. As this is a preliminary working program, only consonants are considered in this project. The oral stop series in Walbiri are not really voiced but typically devoiced, and since there is only the one series it is of little importance that they are written as voiced or voiceless. Hale writes them as voiceless in his dictionary. Gupapuyngu does have a voicing distinction, or rather a fortis-lenis distinction as several linguists claim, but

since it is one of the few Pama-Nyungan languages that has it, it is presumed for the simplification of this project that the lenis series corresponds in a systematic way with the single obstruent series of the proto Gupapuyngu - Walbiri languages and that the omission of the feature [tense] is therefore justified.

Gupapuyngu

b	d	d	ḁ	dy	g
p	t	t	ṭ	ty	k ?
m	n	n	ṇ	ny	ŋ
		l	ḷ		
		ɾ	r	y	w

Walbiri

b	d	dy	ḁ	g
m	n	ny	ṇ	ŋ
	l	ly	ḷ	
	ɾ			
	y	ɾ	w	

Phoneme Inventories for Gupapuyngu² and Walbiri³

Table II

Since most Australian languages possess alveolar, alveopalatal palatalized and retroflexed stops there is no reason to expect that the proto language did not or that the distinctions should be conflated. According to Capell (1962:3), the glottal stop, present in Gupapuyngu, is rare in Australian languages and restricted in its geographical distribution, but since there are several possible sources for this stop it cannot be subsumed under the *k* for example.

This program is not, strictly speaking, dependent on distinc-

² Adapted from S. A. Wurm, *Languages of Australia and Tasmania*, p. 51.

³ Adapted from A. Capell, *Some linguistic types in Australia*, p. 17.

tive feature analysis since the distinctive codes of presence or absence of each feature could be replaced by a purely arbitrary system. Such a system would eliminate redundancy and make possible minimal number of features - a definite advantage in some respects. On the other hand, the code would have to be modified when working on different languages or introducing phonetic variants. In spite of the preference for proper and accurate distinctive features some of the value assignments may be decided arbitrarily in favour of simplicity.

The method for searching for possible cognates by comparing dictionary entries makes assumptions not only about phonological structure but also about word shape. Stress generally falls as close as possible to the front of the word (Capell 1962:4). If stress falls on different syllables during the development of the two languages, that is, on syllable *x* in language A and on syllable *y* in language B, different phonological processes, particularly a variation of lenition processes, must be anticipated. Fortunately, because so much is already established about Gupapuyngu and Walbiri it can be assumed that stress has always been on the first syllable. Tone, too, would have different effects on the phonological development but Capell states (1962:4) that aside from a few languages with 'ornamental' tone, it does not occur in Australian languages.

Another major complicating factor is affixing. Since the program assumes the first syllable (at least) to be the root, the existence of prefixes would pose grave problems. Although Wurm (1972) says that Pama-Nyungan languages are not normally prefixing, it is still possible that some rare forms do have a frozen prefix and this program would not be able to identify the cognate pair if the prefix were only in one of the languages. Nevertheless, any language sufficiently analysed for a dictionary to be produced

would probably have most of the prefixes noted.

There are a number of problems arising from the features used to describe the phonemes of the languages to be compared. When 11 features are used, the shared feature scores for nasals shows /n^y,n/ -7; /ŋ,n/ -8; /ŋ,n^y/ -8. Wurm points out that there is a frequent interchange of /n/, /ny/, and /ŋ/ among related Australian languages. Without modification this program will not select cognate words differing solely by /n^y,n/, without the 'value' level of shared features being so low as to be useless. For example, /L/ and /θ/ also share 7 out of 11 as well as many other unlikely pairs. The correspondence of /g/ and /w/ is also noted by Wurm (1972) but in this analysis they share only 6 out of 11 features.

In order for this program to examine the dictionary entries they must be filed on a computer disc or tape storage device. The format of the entry does not matter as long as it is consistent, although both dictionaries would not necessarily have to be in the same format. The format used here is the same as that of O'Grady's Australian files. Each entry is limited to a single line of length 80 characters, the last 52 of which are reserved for the gloss - sometimes not enough to duplicate the full entry. Where some of the entry is omitted, this is indicated by suspension points. A line looks like this:

Language	C ₁	V ₁	C _{2a}	C _{2b}	C _{2c}	Etcetera	Gloss
----------	----------------	----------------	-----------------	-----------------	-----------------	----------	-------

GUP		D	O	P		ULU	Gambling
-----	--	---	---	---	--	-----	----------

The first three characters name the language. Each segment position of the initial grouping is reserved two spaces. This is because of the impracticality of using diacritics and special symbols with the computer. Even when there is no consonant at C_{2b} or C_{2c}, the space is left empty so as to separate the supposed root from other morphological material. Because of the usual lack of a reconstructible

third syllable in Pama-Nyungan roots, all but the initial vowel of the 'Etcetera' group is normally irrelevant for the reconstruction of the root; hence the original spelling found in the dictionary was kept rather than modifying it as was done for the initial grouping.

The orthographic representations of each segment are chosen so as to simplify working with a computer. It is possible to use all the normal phonetic symbols but because of the increased complexity of the programming involved in doing so, it is better to be content with the much simpler system of allowing two character spaces per phoneme. In the orthography used for this computer program, alveopalatals /t^y n^y l^y/ are written TY, NY, LY; /ŋ/ as NG; dentals /t d n/ as TH, DH, NH; retroflexed consonants /ɖ ɗ ɳ ɽ/ as RT, RD, RD, RL, /r/ as RR; and all others as a single capital letter plus an empty space. Some sounds may have varying spellings in the different writing systems of different languages. This approach is a great advantage as it allows dictionaries using any system of spelling to be used. Each spelling is defined in the array in terms of distinctive feature values such that *b* and *p* have the same specifications if they represent the same phoneme and different specifications if they represent different phonemes.

This program for extracting possible cognates considers only the phonetic form (actually the spelling), and ignores the meaning. After the pairing process the glosses of the two entries must be compared. If the semantic connection seems plausible, then the pair may tentatively be considered cognate. When a number of putative cognates have been found exhibiting the same systematic sound correspondence, then the weight of numbers may be taken as support for cognation and for the idea that the sound correspondences are systematic. Added to the relatively concrete and technical problem

of quantifying the relation between phonemes is the abstract and more difficult task of deciding what is a plausible semantic connection. The pair

WAL R AMPAKU light in weight

GUP RDAMBA light in weight

are eminently acceptable as cognates. On the other hand, the pair

WAL RRANJARR-KA to carry a full load

GUP RDANDJA inside (hold of ship)

are similar enough to arouse interest yet far from convincing.

O'Grady (n.d.) gives a putative cognate set in which the Gupapuyngu and Walbiri forms mean 'dry, dried up, burnt, stale, overcooked' and 1. NEGATIVE 2. 'absent', respectively while the form in Pintupi means 'continually, still'. Were it not for the close similarity of the phonetic form (/rawak/, /lawa/, /rawa/, respectively), and without independent semantic corroboration from other roots, it would be extremely difficult to establish cognation with such divergent meanings. It is only by combining the phonological and semantic evidence that relationships can be stated with a reasonable degree of confidence. When several clear cases have been found, the systematic sound correspondences exhibited can add credence to other pairs with less obvious semantic connection. Without an efficient way of codifying the semantics that includes all possible types of shift a search of this type must begin with the phonological shape.

The problems of writing this program are greatly simplified in a number of ways. First, the dictionary sections used contain words whose initial consonants are already known to correspond in a systematic way between the two languages. Consequently, the compounding of problems involved in comparing more than one thing at a time is avoided. Similarly, only the first consonant of the

cluster following the first vowel is examined. In the plausible case of a /ŋk/ - /g/ correspondence only the /ŋ/ of the cluster is compared with the /g/ where both segments ought to be taken into consideration. Vowels are ignored. This is perhaps not too serious an omission since most Australian languages have basically only /i a u/ (Capell 1962) and with primary stress on the first syllable the vowel would normally be expected to be relatively stable. The program is already confronted with many problems in a pair of Australian languages where initial dropping has led to the development of monosyllabic roots in which initial clusters are found and it can only be expected that comparing languages with more complicated word shapes would present even more difficulties.

As this program considers only the first segment of the root-medial consonantism it cannot be very selective. In this preliminary trial the compared dictionary lists have as initials phonemes that are known to correspond so that a necessary step would be to compare the initials in case of unknown correspondence. The vowels, too, need to be considered so that in the case of the present print-out many of the improbably cognate pairings such as

WAL RRAKU hole in the ground

GUP RDIKADIKA curly

would be avoided. Another less easily solvable problem is that of the proto form having as reflexes a single segment in one language and a cluster in the other. When more than one position in the word is considered, some kind of averaging or dependency must be worked out.

While this preliminary program is limited in scope and is subsequently not as selective as one would like, it does show potential for being a useful tool. There are possibilities for improvement by choosing different distinctive features or by weighting

those considered more relevant. Given the typical complexity of language it is probably impossible to refine such a program to the point of its selecting all and only those pairs which are cognate. Nevertheless, the results of this first attempt indicate that a system based on the principles used here could be useful as an initial step in isolating pairs of words which have high potential as cognates from the thousands of words which are unrelated.

REFERENCES

- Capell, A. 1962. *Some Linguistic Types in Australia*. Oceania Linguistic Monograph No. 7. University of Sidney.
- Hyman, Larry H. 1975. *Phonology: Theory and Analysis*. New York: Rinehart & Winston.
- Wrum, S. A. 1972. *Languages of Tasmania and Australia*. The Hague: Mouton
- Unpublished works
- Hale, Kenneth. 1974. *Warlpiri-English Vocabulary*. M.I.T.
- Lawton, D. W. and Beulah Lowe. n.d. *A Temporary Gupapuyngu Dictionary*. Typescript.
- O'Grady, G. N. n.d. *Change in Australian Languages: Pama-Nyungan Comparative REconstruction*. Mimeograph.