Numerical Taxonomy and the Classification of Salish Indian Languages

David B. Kronenfeld
Lynn L. Thomas
Department of Anthropology
University of California
Riverside, California 92502

Prepared for presentation at 9th International Conference on Salishan Languages, University of British Columbia, 12-14 August 1974.

[Note: This is an extremely rough and extremely tentative first draft. To protect yourself, you should consult with the authors before citing anything in this paper since we are reserving the right to change our opinion on anything stated herein. We should also point out that neither of us is a specialist on either Northwest Coast languages or cultures. We are primarily addressing ourselves to a methodological problem; we are using Swadesh's data, and so our substantive results can only be as good as his data (although, of course, our results could be worse than his data--but we hope not.)]

I The Problem

In 1950, Morris Swadesh presented a classification of Salishan languages; his classification was based on percentage similarities in 'selected basic vocabulary' of all pairs of thirty Salishan languages and dialects (see matrix of similarities, Table 1). In 1962, Dyen proposed a general procedure for classifying languages into family trees based on the same kind of data as Swadesh used; Dyen in fact exemplified the method with the use of Swadesh's matrix of similarities for the thirty Salishan languages. In 1969, Jorgensen using a different procedure--one based on the work of numerical taxonomists in biology--provided a third classification of the same thirty languages. Swadesh, Dyen, and Jorgensen all realized that the data used--the matrix of similarities--reflected for the most part two distinct kinds of historical processes: genetic drift and borrowing. While Swadesh and Jorgensen did not distinguish between the effects of drift and borrowing in their classifications, rather letting them reflect the effects of both, Dyen did attempt to eliminate particular similarities values which were judged to be 'inflated' by borrowing. It is our purpose in this paper to provide yet a fourth classification of the same matrix of similarities of Salishan languages, one which we think an improvement on the earliest classifications, especially in distinguishing the effects of borrowing from the effects of drift. Because of doubts about the adequacy of the Salish data, we are not so much attempting to definitively classify the Salishan languages (see Dyen 1962: 256, footnote 7 on Swadesh's doubts about the data) as to present a method. We will outline the problem first, by briefly discussing the classifications

and classification methods used by Swadesh, Dyen, and Jorgensen. Second, we will describe our method and the assumptions on which it is based. Third, we will describe the major differences in classifying Salishan languages that our method yields as compared with the results of the earlier methods; we will give our reasons for thinking our classification preferable to the earlier ones. Finally, we will indicate what further work needs to be done in the way of improving on the method and testing it. Throughout our discussion we will be making use of special and novel assumptions about the nature of the data and the workings of our method. Some of these assumptions are certainly subject to debate. However, insofar as they place strong constraints on the data, and insofar as these strong constraints are met, we infer some support for the assumptions.

It was Swadesh's purpose in his 1950 paper to determine "...the degrees of relationship or the stretch of time needed to account for the formation of [the Salishan] stock" (Swadesh 1950:157). To this end he applied the assumption and formulae of what came to be called 'glottochronology' to his similarities matrix. His results included a classification and a set of accompanying 'indicated time depth periods'. Like Dyen and Jorgensen, we will not rely on any method aimed at discerning time depth because of the serious problems of inferring time depth; we will discuss only Swadesh's classification. This is possible because the 'indicated time depth periods' and the 'percent of common vocabulary' are monotonically (inversely) related, giving the result that the classification can be viewed in terms of either. In this paper in other words, we will be doing 'lexicostatistics' but not 'glottochronology'. Although Swadesh does give numerical criteria--expressed as 'units of linguistic distance'--for his clusterings into 'languages',

and language 'groups', 'branches', and 'divisions', his classification method seems to have been largely impressionistic. What we mean by 'impressionistic' can be seen from the following hypothetical example.

Imagine that three languages L_1, L_2, and L_3 have been clustered into a group 'A' on the basis of having higher similarity values among themselves than with any other language. Imagine further that two other languages, L_4 and L_5, are candidates for inclusion into a grouping B which includes grouping A. If genetic drift and no other factors were operating in a very idealized world, we might expect the percent of common vocabulary values for the pairs L_1-L4, L2-L4, and L3-L4 to all be the same as would the values L1-L5, L2-L5, L3-L5 all be the same. There would then be just two values to consider, the value for L4 with A and the value for L5 with A; we simply choose the larger value and include the language associated with that value in the new B clustering. The non-included language would either be included in a later grouping with the languages of A and B, presumably reflecting an earlier common ancestor or more rapid divergence from the ancestor, or it would go in some other grouping. But we know that the similarities data are not so simple and that borrowing, and other factors, offset the simple effects of drift. Assume, then, that the percent of common vocabulary values for our hypothetical example are as follows:

	Group A		
	L_1	L_2	L_3
L_4	58	57	75
L_5	64	69	62

Given such a situation, what criterion might Swadesh have used in deciding whether L_4 or L_5 is the more closely related to A? Among, the criteria applicable to this situation are the following possibilities:

1) Choose the language (L_4 or L_5) with the largest average (or some other measure of central tendency) value with the languages of cluster A.

By this rule, L_5 would be selected.

2) Choose the language with the largest single value with the languages of cluster A.

By this rule, L_4 would be selected on the basis of its value with L_3.

3) Choose the language with the largest minimum value with any of the languages of cluster A.

By this rule, L_5 would again be selected, on the basis of its value with L_3. We do not know which of these criteria Swadesh used in developing his classification. He appears to have used some kind of central tendency criterion in many instances. However, he does give one important basis for deciding which kind of rule to follow in some instances.

Swadesh says that if two languages have been in close contact for a long time, they will have influenced each other either with the effect that both languages will have changed less rapidly than other languages which did not have a similar contact situation, or with the effect that the two languages will have changed more rapidly than other languages; in both events the two languages will trend in the same direction. Under such conditions Swadesh suggests using a minimum similarities value (rule 3 above, or perhaps rule 3 combined with rule 1), thereby not allowing the increase in similarity of pairs of languages in close contact to bias the classification away from a 'pure' genetic one. We doubt that Swadesh used this insight in his classification in any systematic way; both the

classification and his discussion of it appear to mix the effects of drift and borrowing. But in our proposed method the insight is fully applied.

Swadesh's classification gives four divisions of Salish: Bella Coola, off to the north and separated from the other Salishan languages; the Coast Division with five branches (North Georgia, South Georgia, Puget Sound, Hood Canal, and Olympic); the Oregon Division to the south (represented by Tillamook) and separated from the others; the Interior division, east of the Cascades. The full classification is given in Table 2 (with Swadesh's abbreviations for the languages.

Unlike Swadesh, Dyen gives a fully explicit procedure for deriving a family tree from a matrix of similarities. We need not give a complete presentation of his method but will instead focus on its two aspects that are most pertinent to comparisons with Swadesh's, Jorgensen's, and our methods. Both of these aspects are indicated in the following rule Dyen developed:

> If languages A,B...N constitute a group, their respective percentages with the same non-member are averaged, except those that are demonstrably distorted (Dyen 1962:156).

First, Dyen is using averages (rule 1 above) rather than maxima or minima in figuring the similarities values of non-members with members of previously formed clusters. Second, he eliminates from the averaging percentages which are judged to be inflated by borrowing [the 'distorted' percentages in the above passage]; "...borrowing is a possible explanation of the fact that two members show significantly different percentages with the same non-member" (Dyen 1962:156). As with Swadesh's insight on the use of minima mentioned above; it is presumed in Dyen's method that there is no large systematic force which would decrease similarity values in the way

borrowing would increase them. It remains only to note that Dyen uses a rule-of-thumb value of 9.5% as a lower level of 'significant difference' between two similarities values. Dyen's resulting classification is quite similar to Swadesh's, but with the following differences which Dyen considers substantial enough to support his classification over Swadesh's (see Table 3). First, Swadesh has his Lkungen Group (Lm, Lk, Cl) with Squamish and the Nanaimo Group (Fr, Nn) in his South Georgia Branch (with Nt); Dyen places his Lkungen Branch (Lm, Lk, Cl) separate from his South Georgia Branch (Fr, Nn, Sq, Nt). Dyen thinks that Swadesh's results were biased by inflated percentages of Lk-Fr (53%) and Lk-Nn (54%); these three languages (Lk, Fr, Nn) are contiguous spatially on the eastern end of Vancouver Island.

Second, Swadesh combines his Satsop Group (Cw, Ch, Sa) with Lo and Qu into an Olympic branch; Dyen keeps these two clusterings apart "...because their common highest percentage, 43%, is not significantly different from the Satsop Branch's next highest percentage 38% with Twana, this being Twana's highest percentage" (Dyen 1962:161). Third, Swadesh treats Li, the Thompson Group, the Okanagon Group, Cm, and Cr as coordinate members of the Interior Division while Dyen places Cm and Cr with Sp, Ka, Pe, and Ok (Columbia Branch) and Li with Th and Sh (Lillooet Branch); since the latter clustering is in spite of inflated percentages of Cr-Ka, Cr-Pe, Cr-Sp, it can be surmised that Dyen's mode of averaging values and his use of the 'significant difference' rule are what caused the clustering which yields the Columbia Branch. Since Dyen finds so few values inflated by presumed borrowing and since, in the case of the Interior languages,

clusterings are made in spite of inflated percentages, it would appear that his averaging and the 'significant difference' rule account for the bulk of the differences from Swadesh; without the 'significant difference' rule, Dyen's clusterings of the Coast languages would be practically identical with Swadesh's.

Jorgensen briefly describes his method in these terms:

> Treelike diagrams are determined by a nonmetric technique for finding the smallest euclidian space for several points.... Briefly, all the unit pairs in the sample are scanned and the pair with the highest, i.e., closest, coefficient is joined. Then the distance between the nearest actual number of each pair is measured to the centroid of the other pair and attached at the center of gravity....
>
> (Jorgensen 1969:123)

Although we have not examined the computer program Jorgensen used, we can guess from the above statements and from Jorgensen's classification that he used an averaging method. Jorgensen's results are given in Figure 1. Note that his results are different from Swadesh's and Dyen's in having many more 'nodes', or points of divergence. This makes Jorgensen's results and those of Swadesh and Dyen somewhat difficult to compare; but while Swadesh's and Dyen's schemes describe the data less fully, Jorgensen's scheme, especially at the higher levels, can be collapsed to give results comparable to the other two classifications. Jorgensen, like Swadesh, has not attempted to differentiate genetic drift from borrowing in his classification [although he does treat borrowing separately, by means of a 'contact interval' analysis technique devised by Elmendorf 1965]; however, despite his disavowal, his use of tree diagrams with one-many mappings does indicate genetic-type as opposed to borrowing relations.

Jorgensen's procedure retains the same four divisions of Swadesh and Dyen: Coast, Interior, Bella Coola, and Tillamook. Jorgensen's classification agrees with Swadesh's in the South Georgia Branch clustering (including Lkungen Group with Sq, Fr, Nn, and Nt) and in the Olympic Branch clustering (Lo, Qu, Sa, Ch, and Cw); Jorgensen agrees with Dyen in the Lillooet Branch, Columbia Branch clusterings within the Interior Division.

II Our Method

There are two kinds of tasks that one could ask of a clustering technique. One is to produce a configuration that as accurately as possible summarizes (or can reproduce) the inputted similarity data; a second task is to produce a configuration that represents as accurately as possible some posited underlying reality in a situation in which that reality (at least in part) produced the inputted similarities.

The first task is a purely formal one, and is relatively self-contained in the sense that one can compare the output of the clustering technique with the inputted data and without external information immediately, determine how well the clustering technique did its job. Success at this task in no way depends on the content of the specific similarities (as opposed to the relative sizes of the numbers) or of the reality which they are supposed to represent.

The second task differs in several respects from the first. First, it is not self-contained; information beyond the inputted similarities is required for the evaluation of its success. This extra information will be precisely the information that is being sought--which is to say that in practice there will be no possible direct test of the success of the technique;

one will have to rely on indirect tests involving the consistency of this solution with other information that one has about the posited underlying reality. One can devise simulations to directly evaluate the adequacy of techniques for the performing of this task, but such evaluations are only as good as the simulations are insightful and complete.

One should note that a best solution to the first task is not necessarily best for the second. The two tasks are equivalent only when the particular underlying reality is the only source of the similarity data, and when errors (whether of measurement or of the process by which the underlying reality produces the similarity data) are randomly distributed. If errors are considered to be non-random, then one would like ones technique to take account of the error bias. If several different underlying realities combine to produce the similarity data, then one would like one's technique to filter out the separate effects of each, or at least to isolate the effects of the particular source of similarity that one is primarily concerned with. In any of these cases, the best representation of a particular underlying reality will be different from the best direct representation of the inputted similarity data because one will be assuming that the inputted data is biased and because one will be trying to take account of that bias. These considerations of error and of variables other than the one being studied also preclude any general solution to this task, even for a single clustering technique--that is, the best solution will vary from one empirical problem to another, depending on the shape of other underlying variables, and the kinds of error bias assumed.

Jorgensen's procedure retains the same four divisions of Swadesh and Dyen: Coast, Interior, Bella Coola, and Tillamook. Jorgensen's classification agrees with Swadesh's in the South Georgia Branch clustering (including Lkungen Group with Sq, Fr, Nn, and Nt) and in the Olympic Branch clustering (Lo, Qu, Sa, Ch, and Cw); Jorgensen agrees with Dyen in the Lillooet Branch, Columbia Branch clusterings within the Interior Division.

II Our Method

There are two kinds of tasks that one could ask of a clustering technique. One is to produce a configuration that as accurately as possible summarizes (or can reproduce) the inputted similarity data; a second task is to produce a configuration that represents as accurately as possible some posited underlying reality in a situation in which that reality (at least in part) produced the inputted similarities.

The first task is a purely formal one, and is relatively self-contained in the sense that one can compare the output of the clustering technique with the inputted data and without external information immediately, determine how well the clustering technique did its job. Success at this task in no way depends on the content of the specific similarities (as opposed to the relative sizes of the numbers) or of the reality which they are supposed to represent.

The second task differs in several respects from the first. First, it is not self-contained; information beyond the inputted similarities is required for the evaluation of its success. This extra information will be precisely the information that is being sought--which is to say that in practice there will be no possible direct test of the success of the technique;

one will have to rely on indirect tests involving the consistency of this solution with other information that one has about the posited underlying reality. One can devise simulations to directly evaluate the adequacy of techniques for the performing of this task, but such evaluations are only as good as the simulations are insightful and complete.

One should note that a best solution to the first task is not necessarily best for the second. The two tasks are equivalent only when the particular underlying reality is the only source of the similarity data, and when errors (whether of measurement or of the process by which the underlying reality produces the similarity data) are randomly distributed. If errors are considered to be non-random, then one would like ones technique to take account of the error bias. If several different underlying realities combine to produce the similarity data, then one would like one's technique to filter out the separate effects of each, or at least to isolate the effects of the particular source of similarity that one is primarily concerned with. In any of these cases, the best representation of a particular underlying reality will be different from the best direct representation of the inputted similarity data because one will be assuming that the inputted data is biased and because one will be trying to take account of that bias. These considerations of error and of variables other than the one being studied also preclude any general solution to this task, even for a single clustering technique--that is, the best solution will vary from one empirical problem to another, depending on the shape of other underlying variables, and the kinds of error bias assumed.

In this paper we are concerned with relations among languages; in particular we are concerned with the problem of sub-grouping--i.e. the problem of determining the precise interrelations of a set of languages that are already known to be genetically related to one another. The similarity data that we are using consist of percentages of shared basic vocabulary (calculated by Swadesh, as explained above). Two languages can share an item of vocabulary a) by both inheriting it from a common ancestor (where ever that ancestor got it from), b) by one borrowing it from the other, or both borrowing it from some third source, c) by means of some sort of universal process of sound symbolism, or d) by chance. In this paper we are assuming that the effects of c and d are small enough to ignore (except as a residue). We are also assuming that at any given time depth a language has one and only one parent, and thus that the family tree of any group of languages will map from a unique beginner (proto-language) through a series of one-to-many nodes (other later proto-languages) to the set of existing languages. We make this assumption of a taxonomic kind of structure a) because in order to select an appropriate kind of clustering technique we have to know what kinds of clusters we are looking for, and b) because this is the kind of structure traditionally posited in historical linguistics.

Different kinds of clustering programs are sensitive to different kinds of structures. For example, multi-dimensional scaling can provide an n-dimensional spatial representation (for a smallest n) of the relations among a number of points, but its constraints break down as the number of points approaches the number of dimensions. Multi-dimensional scaling,

thus, is useful for finding paradigmatic type structures (in which a small number of dimensions intersect to produce a large number of points and in which all dimensions are relevant to all points), but is incapable of finding taxonomic type structures (in which the number of dimensions or distinctive features is only slightly less than the number of points, and in which each dimension is only relevant to one node--and thus only to the points dominated by that node). Hierarchical clustering, on the other hand is particularly sensitive to taxonomic type structures, and incapable of finding truly paradigmatic structures.

In the case of the present problem, we will use a hierarchical clustering technique to find the genetic (family tree) relations among the Salish languages; we will assume that the remaining basic vocabulary similarities not accounted for by this tree are the result of the other three sources of similarity that we describe earlier, of borrowing in particular. We will, then, look at the history and geography of the language communities in question in order to see how likely such borrowing is and to see how unlikely it is that borrowing provides an effective alternative theory to our postulated genetic relations. Such an examination will necessarily involve a few assumptions concerning the conditions under which vocabulary (especially "basic vocabulary") borrowing takes place; at the appropriate time we will state what our assumptions are.

We want, then, a hierarchical clustering technique that is most likely to give us a true picture of the genetic relations among Salish languages. Such a technique will not be the technique which best reproduces the inputted similarity data since we are assuming that that data is the product of (non-taxonomically structured) borrowing as well as of the genetic

In this paper we are concerned with relations among languages; in particular we are concerned with the problem of sub-grouping--i.e. the problem of determining the precise interrelations of a set of languages that are already known to be genetically related to one another. The similarity data that we are using consist of percentages of shared basic vocabulary (calculated by Swadesh, as explained above). Two languages can share an item of vocabulary a) by both inheriting it from a common ancestor (where ever that ancestor got it from), b) by one borrowing it from the other, or both borrowing it from some third source, c) by means of some sort of universal process of sound symbolism, or d) by chance. In this paper we are assuming that the effects of c and d are small enough to ignore (except as a residue). We are also assuming that at any given time depth a language has one and only one parent, and thus that the family tree of any group of languages will map from a unique beginner (proto-language) through a series of one-to-many nodes (other later proto-languages) to the set of existing languages. We make this assumption of a taxonomic kind of structure a) because in order to select an appropriate kind of clustering technique we have to know what kinds of clusters we are looking for, and b) because this is the kind of structure traditionally posited in historical linguistics.

Different kinds of clustering programs are sensitive to different kinds of structures. For example, multi-dimensional scaling can provide an n-dimensional spatial representation (for a smallest n) of the relations among a number of points, but its constraints break down as the number of points approaches the number of dimensions. Multi-dimensional scaling,

thus, is useful for finding paradigmatic type structures (in which a small number of dimensions intersect to produce a large number of points and in which all dimensions are relevant to all points), but is incapable of finding taxonomic type structures (in which the number of dimensions or distinctive features is only slightly less than the number of points, and in which each dimension is only relevant to one node--and thus only to the points dominated by that node). Hierarchical clustering, on the other hand is particularly sensitive to taxonomic type structures and incapable of finding truly paradigmatic structures.

In the case of the present problem, we will use a hierarchical clustering technique to find the genetic (family tree) relations among the Salish languages; we will assume that the remaining basic vocabulary similarities not accounted for by this tree are the result of the other three sources of similarity that we describe earlier, of borrowing in particular. We will, then, look at the history and geography of the language communities in question in order to see how likely such borrowing is and to see how unlikely it is that borrowing provides an effective alternative theory to our postulated genetic relations. Such an examination will necessarily involve a few assumptions concerning the conditions under which vocabulary (especially "basic vocabulary") borrowing takes place; at the appropriate time we will state what our assumptions are.

We want, then, a hierarchical clustering technique that is most likely to give us a true picture of the genetic relations among Salish languages. Such a technique will not be the technique which best reproduces the inputted similarity data since we are assuming that that data is the product of (non-taxonomically structured) borrowing as well as of the genetic

relations obtaining among the languages. We will need independent means of evaluating the fit. We want a technique that will filter out the effects of borrowing. In order to find such a technique we must first ascertain the effects that borrowing would have on a table of similarities that otherwise only reflect genetic relations. We can then look for a method which is least influenced by these specific effects.

With the preceding goal in mind, now let us turn to the Johnson Hierarchical Clustering program (Johnson 1967). This program constructs trees by two different methods. Both methods use as data a similarity matrix in which larger numbers represent greater similarity (correlation coefficients, percentage of shared vocabulary, or etc.). The methods work down from the highest similarity level in the table to the lowest. The items which are similar to one another at the highest similarity that level are joined together into a cluster; that cluster becomes a unit for subsequent clustering. Both methods then go to the next highest similarity joining those items into a cluster (either forming a new cluster, joining new items into existing clusters, or joining two or more existing clusters into a new, larger cluster) which also becomes a unit for subsequent clustering, and then to the next highest similarity, and so forth until all the separate items have been joined into one single cluster. The program indicates in its output the levels of similarity at which the successively more inclusive clusters were constructed. The two methods differ in the criteria by which new items get joined into existing clusters and by which existing clusters get joined into larger clusters. In the first method, the <u>minimum</u> or connectedness method, a new item is joined into an existing cluster at the highest similarity level at which

it is similar to at least one item already in the cluster. Similarly, two clusters get joined together at the highest level at which at least one item from one cluster is similar to at least one item in the other cluster. The minimum method constructs its clusters according to the highest similarity level of items in exiting clusters to other items. In the other method, the maximum or diameter method, a new item get joined into an existing cluster at the highest similarity level at which it is similar (at that level or a higher level) to all items already in the cluster. Similarly, two clusters get joined together at the highest level at which all items in the one cluster are similar (at that level or a higher level) to all items in the new cluster. The maximum method does not join a new item into an existing cluster until it reaches a similarity level at which the item is similar to all items already in the cluster, i.e. until the level of the new item's lowest similarity to an item already in the cluster. Other methods exist (though not used by the Johnson program) which join new items to existing clusters according to the level of one or another kind of average value of the similarities of new items to items already in the cluster (cf. Sokol and Sneath 1966). The reader is asked to keep these abstract descriptions in mind for the moment; we will later examine some concrete examples of the use of the minimum and maximum methods.

Let us compare the two methods. If the imputted data is produced by a true taxonomy and if the similarity data accurately represents that taxonomy, then the two methods of the Johnson program should produce exactly the same results; the highest similarity value of a new item with any item in an existing cluster should be the same as the lowest

similarity value of the new item with any itme already in the cluster. In other words, if the cluster already contains horses, cows, foxes and tigers and if the new item is red snapper, then both methods should join red snapper in at exactly the same level since the taxonomically relevant similarity is between fish and mammals, and since a fish is exactly as similar to one mammal as to any other mammal ("similar to" means "closely related to" in this case). Similarly, a cluster containing red snapper, trout, and flounder would be joined by both methods to our horse, cow, fox, tiger cluster at exactly the same point as was red snapper alone--for the same reason. Figure A below illustrates this situation.

Fig A

Similarity Matrix:

	Hor	Cow	Fox	Tig	Tro	RSn	Flo
Horse							
Cow	H						
Fox	M	M					
Tiger	M	M	H				
Trout	L	L	L	L			
Red Snapper	L	L	L	L	M		
Flounder	L	L	L	L	M	M	

High (H), medium (M), and low (L) indicate relative similarity levels.

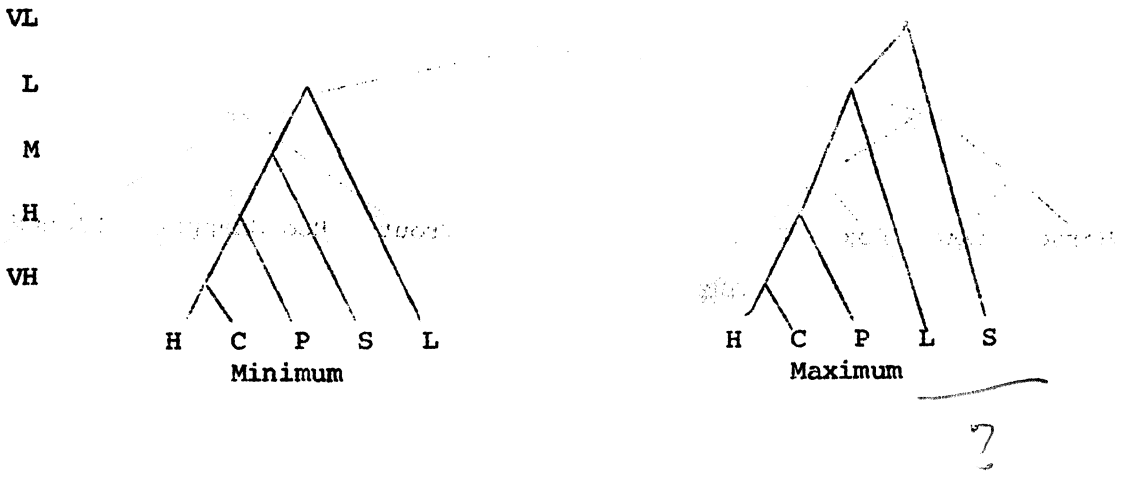
Hierarchical Clustering Tree (by both min and max):

The two methods produce different results when the similarity matrix diverges from a direct representation of a true taxonomy. To see how they differ, let us take another animal example. Our existing cluster contains horses, cows, and porpoises. Our new items are sharks and lizards. Our similarity measure, this time, is less than a perfect reflection of true taxonomic relations, and so is in part influenced by the fact that sharks and porpoises look a lot alike and live at least a little bit alike. In Figure B below we can see that, in this highly oversimplified example, the minimum method puts sharks into the cluster before lizards because of the high porpoise--shark similarity. The maximum method puts lizards in before sharks because lizard's similarity to everything in the cluster is L, while shark's lowest level of similarity to items in the cluster is VL (very low) to horse and cow. In linguistic applications, borrowing is the logical counterpart of the convergence of sharks and porpoises.

Fig B

	H	C	P	L	S
Horse					
Cow	VH				
Porpoise	H	H			
Lizard	L	L	L		
Shark	VL	VL	M	VL	

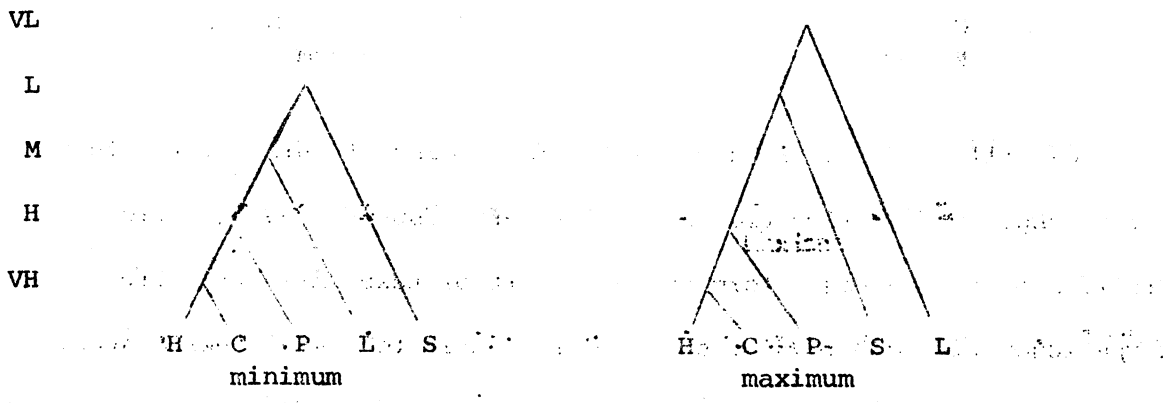
VH: very high
H: high
M: medium
L: low
VL: very low



In spite of this example, however, the maximum method is not always better for all purposes than the minimum method. Let us suppose that our similarity matrix does correctly represent the relation of sharks to porpoises, but falls into another error by taking the relationship between the features of warm blood and sea living too seriously and therefore greatly undesestimates the lizard-porpoise similarity. In Figure C we can see that in this case the minimum method gives us the better representation of the underlying reality, while the maximum method is most affected by the error.

Fig C

	H	C	P	L	S
Horse					
Cow	VH				
Porpoise	H	H			
Lizard	M	M	VL		
Shark	L	L	L	L	

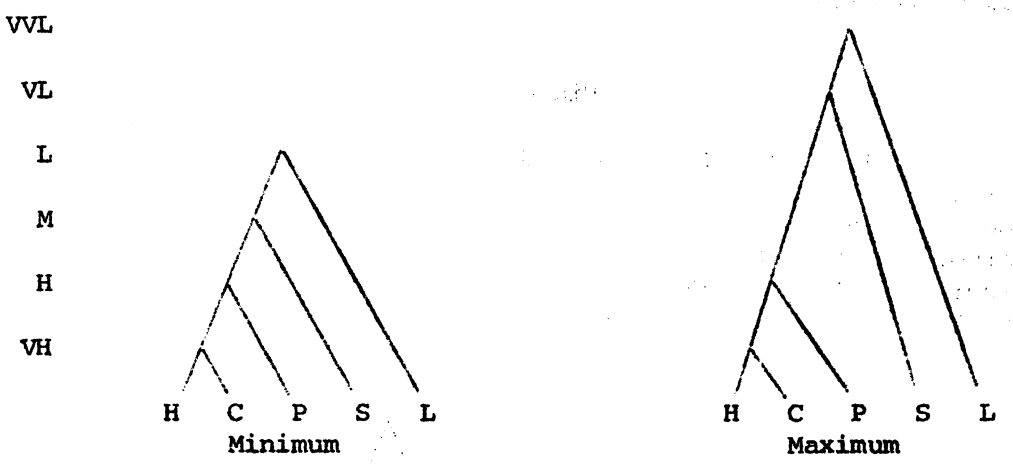


In the presence of both errors the underlying reality may simply be unrecoverable, as we can see in Figure D.

In the presence of both errors the underlying reality may simply be unrecoverable, as we can see in Figure D.

Fig D

	H	C	P	L	S
Horse					
Cow	VH				
Porpoise	H	H			
Lizard	L	L	VVL		
Shark	VL	VL	M	VL	



These illustrations lead us to the main point of this paper. In our Salish Language classification problem we are assuming that genetic relations among languages form the same sort of taxonomic structure as do the phylogenetic relations of species of animals to one another. We are also assuming that the only other factor with a major influence on the similarity values is the borrowing of basic vocabulary by one language from another or by both from some third language. If we can show that borrowing is more likely to produce one kind of deviation from the similarities produced by the genetic relations (cf Fig B) than it is the

other kind of deviation (cf Fig C), then we will be able to use one of the methods as a reasonably direct representation of the genetic relations among the languages rather than having to guess at the different effects of shared descent and borrowing.

In the discussion that follows we will show that borrowing is much more likely to raise specific pairwise similarity scores than to lower them, and thus that the maximum method gives a much more accurate picture of the genetic relations among a group of languages than does the minimum method or any obvious kind of averaging method. In this situation the minimum method structure represents the effects of borrowing on the pattern relations. We will first show this result in a general way with some simulated examples. We will then apply our technique to the Salish languages and adduce some independent ethnographic and geographical data in support of our subgrouping and the history of separations and contacts that it entails.

In this discussion we are speaking of a word as related to another word if it has the kind of sound meaning correspondence spoken of by Swadesh in his lexicostatistics discussions. Pairs of words lacking such correspondences are considered unrelated--even if each member of a given pair is cognate to some other word in the language of the other word of the pair. This kind of relatedness can only take on one of two values for any given pair of words: "related" or "unrelated"; at the level of individual words there are no <u>degrees</u> of relatedness. There are degrees of relatedness among languages; the degree of relatedness between two languages is represented in the data set by the percentage of basic vocabulary words in the two languages which are "related". It is important

to note the "all or nothing" nature of relatedness between pairs of words in Swadesh's system because several of the inferences which we will draw in the discussion that follows concerning the effects of different kinds of borrowing will depend on this fact.

Now let us turn to Figure G. Language D's pre-replacement vocabulary will consist of four parts. a) words unrelated to the family by whatever rule is used for calculations shared as in vocabulary, b) words related to a branch of the family that does not include E, c) words related to a higher branch of the family which includes both D and E (which words are also possessed by E), d) words related to a higher branch of the family which includes both D and E (which words, however, are not possessed by E). E's pre-donation vocabulary will include c) from above, as well as e) words related to a higher branch of the family which includes both D and E (which words, however, are not possessed by D), f) words unrelated to the family by whatever role is used for calculating shared basic vocabulary, and g) words related to a branch of the family that does not include D. When D borrows vocabulary from E, any of its four kinds of words may be replaced by any of E's four. To explore the effects such a borrowing might have on the apparent taxonomic tree otherwise (pre-borrowing) implicit in the matrix of basic vocabulary similarities, we now wish to indicate which replacer/replacee combinations can occur and what effect each possible combination will have on the pattern of similarities.

Type a words in D can be replaced with types e, f, and g from E. They cannot be replaced with c because, by the definition of c, D's equivalent word would also have to be a type c one. Type b words in D

But they can be, e.g., dialect borrowing

I should think any word could be replaced by any other word.

can be replaced with types f and g from E. Type c replacers are ruled out as before, and type e are ruled out because b words are defined as belonging to D's branch (vs. E's branch) while e words are defined as belonging to both branches (even if D itself doesn't have them). Type c words can only be replaced by type c words, as explained. Type d words can only be replaced by type f replacers since d are words that both branches have, (even if E itself doesn't have them; since E doesn't have the d word and since there can be no other family possibility (by the definition of d), E can only offer an f word. dialect borrowing?

In the discussion that follows an expression such as f/a will signify that a type f word from E replaces a type a word in D. c/c borrowings will obviously have no effect on the number of cognates between D and E, nor on the number of cognates between D and any other language. f/a borrowings will increase the number of D E similarities, but, since both f words and a words are non-family, will have no affect on D's similarities to any other languages in the family. f/b borrowings will increase the number of D E similarities; these borrowings will also decrease D's similarities to other languages in its branch since words from its branch are being replaced by non-family words. f/d borrowings will increase the number of D E similarities; these borrowings will also lower D's similarities to other languages in the family since family words are being replaced by unrelated words. e/a borrowings replacing unrelated words with family words, will increase D's similarities to the rest of the family. g/a borrowings, by replacing non-family words with words from E's branch, will increase D's similarities with E's branch, but will not affect the D's

similarities to its own branch. g/b borrowings, by replacing D's branch words with E's branch ones, will both increase D's similarity to E's branch and decrease D's similarity to its own branch.

c/c borrowings obviously have no affect on the apparent taxonomic tree embedded in the matrix of percentages of shared cognates since they change none of the percentages. Our technique for inferring a taxonomic tree from a matrix of similarities is not affected by a particular raised similarity value such as the D E value would be because of f/a type borrowing--as long as D and E still have higher similarities to languages in their own branches. Similarily, raising D's similarities to all the rest of the family with e/a borrowings would not affect our inferred taxonomic tree--except at the lowest levels in the unlikely event that D's increased number of cognates made it more similar to each of the various members of a sub-branch than the members were to each other. Similarly, g/a borrowings would not affect the inferred tree unless D's borrowings were sufficiently numerous to make D's similarities to all of E's branch higher than its similarities to any of its own branch. Note that each of the types of borrowings just mentioned can have no affect on any aspect of the tree beyond simply the placement of D itself--that is, the worst that could happen even if the borrowings were sufficiently large (and E already sufficiently taxonomically close to D) to make D more similar to E than to its closest relatives would be to simply mislocate the single language, D, in an otherwise correct tree. "Sufficiently large" seems quite unlikely for basic vocabulary data unless D and E are already quite closely related since Swadesh (1951:13)

reports studies which show that the absolute amount of borrowed basic vocabulary for a given pair of languages will be well under 10%--i.e. a small number; this means that D's similarities to E's branch must already (naturally) be within 10% of its similarities to the closest members of its own branch before borrowings can affect the inferred tree. The most such borrowings can do is cause minor local mistakes; the cannot affect the basic shape of the tree.

f/d borrowings both raise D's similarities to E and lower D's relations to the rest of the family. But, unless this borrowing makes D more similar to E than to anything else in the family, it will have no affect on the shape of the inferred tree. f/b borrowings have slightly greater possibilities for affecting D's placement since they both raise D's similarities to E and lower D's similarities to D's own branch.

For both of the above kinds of borrowing, there will be little affect on the taxonomic tree. The amount of borrowed similarity can only be small, as explained above, and it can only lower D's relatedness to its own close kin. For reasons just explained such borrowing cannot place D closer to E in the tree unless they are already quite genetically close.

g/b borrowings seem to be the only kind that present any real possibility of affecting the inferred taxonomic tree since they both lower D's similarity to its own branch and raise its similarity to E's branch. This raising and lowering doubles the possible magnitude of the effects of borrowing, and allows the possibility that such effects would be non-trivial. But note that still the only actual effect would be to

place D in the wrong branch; the other relationships would be unaffected. Also note that our discussions of borrowing have been based on the magnitude of total borrowing. To assess the actual likelihood that g/b borrowing could be sufficiently large to have any affects we need to consider how large a proportion of the total borrowing could be represented by the g/b type. To the extent that D's and E's branches split apart recently one would expect a high proportion of the vocabulary of each to be common to both branches (types d, d, or e); the recent divergence of D's branch from E's branch would only allow the development of little basic vocabulary that was peculiar to one or the other branch. To the degree that D and E's branches split apart a long time ago, the proportion of branch words should be higher. This last possibility--that there was a very high degree of borrowing between two distantly related languages (in which words cognate to the receiver's closer relatives but not to the lender were replaced with words cognate to the lender's closer relatives--would pose the biggest threat to the method. It is possible that, in such a situation, D could erroneously be assigned to E's branch; but note that, even here the basic shape of the taxonomic tree would be unchanged by D's mispositioning. The only ways that the true pattern of genetic relations could be sufficiently muddied to be irretrievable by a method such as this would be 1) if D was the only (or one of only a very few) member of its (true) branch in the data set (wherein moving D would wipe the branch out) or 2) if a large number of other languages in D's branch had extensively exchanged vocabulary with a large number of languages in E's branch. The second condition could be detected by the absence of clear patterns in the similarity matrix.

Thus, in general the methods being discussed in this section of the paper should be able, by themselves, to induce the correct genetic tree. In the one kind of case where they might not be able to do so, there should be clear and massive evidence of relatively recent widespread intensive contact among languages in the two groups. If one has such an amorphous matrix from which either of two alternative groupings could be inferred, and if the residual similarities in the one case can be accounted for by known extensive contact but in the other case cannot be so accounted for, then one can use this kind of external evidence to resolve the thus circumscribed version of the classification problem created by this borrowing. It should be noted that such extnesive borrowing of basic vocabulary (sufficient to create such a situation) could only occur extremely rarely and only in most unusual contact situations.

Having laid out the reasons for expecting the maximum method to better represent the true genetic relations among a set of languages than the minimum method, we would now like to turn to the example in Figure E in order to illustrate the actual effects of borrowing on the form of genetic trees recovered with hierarchical clustering. Figure F is based on the same true genetic tree as was E. But we have now included a geographical distribution of the various languages and based borrowing effects on this distribution, as explained in the figure. In that example, languages that geographically border on one another but do not belong to the same immediate cluster have had their similarity scores substantially incremented; the one language that has become geographically separated from all other members of the family has had its similarity scores with all other members of the family substantially lowered. The

second similarity matrix in the figure represents the empirically observed pairwise similarities for the family (after this borrowing has had its effects). The two trees represent, respectively, the results of the two Johnson methods applied to this matrix. We can see that both methods place languages B-J correctly, but that only the Maximum Method places language A correctly. MAX has correctly recovered the shape of the tree, while MIN has not. The relative taxonomic depth of different parts of the family tree has been somewhat disrupted, however, by A. Given the obvious nature of the special geographic effects on A, one could now remove A from the table, re-cluster it, and get an improved figure for the taxonomic depth of the original ABCD-EFGHIJ split.

In Figure G we again consider the same "real" family tree as in Figures E and F. But this time we complicate the effects of borrowing by considering the history of the geographical distribution of the members of the family. The family has moved through stages I - VIII; stages VI and VII have had the indicated effects on the similarity scores. In this example, the MIN method misplaces the BCD branch while the MAX method correctly recovers the shape of the "real" tree. Again, the relative taxonomic depth of certain nodes is somewhat distorted by A; and again, re-clustering without A would better represent these depths.

Figure H considers the same data as did G, but with an additional complication: much of the space between A and B, C has been filled by a language from an entirely different family from A-J K, and K has exchanged extensive vocabulary with its neighbors, B and C. Our linguist has only been considering the problem of sub-grouping languages already known to belong to a single family, but in this case he was misled by K's

strong similarities to B and C and tentatively included it in the family. The MIN method seriously misplaces both K and J, while the MAX method once again recovers the "real" tree. MAX does include K on the tree since it has nowhere else to put it but only at a level of basic vocabulary overlap which is well within the realm of chance, i.e. 2%. On the basis of this outcome, one could next remove K from the similarity matrix and, re-cluster; the result would be Figure G (which would confirm the family subgrouping found by MAX in Figure H).

Again starting with the situation described in Figure G, let us assume that in the transition from Stave V to Stage VI the Y/Z and Y/J similarity scores were (rather massively) lowered by 30%. Figure I presents this situation. In this case, both MAX and MIN produce the same, incorrect, tree by misplacing A. If the historical changes are massive enough and biased enough, the original tree can become irretrievable from this kind of data by this kind of method, but note that MIN still is not better than MAX and that the degree of bias and the size of its effect in this example has been quite spectacular. The degree of error in this example may also be considered fairly small in the sense that only the one language is out of place in an otherwise correct tree.

We would like to suggest that the above results look quite good. To give the reader an idea how good, and to indicate that the structures found do indeed inhere in the data matrix rather than tautologically in the method, we would now like to consider one further example. Figure J deals with the same languages that were introduced in Figure G. Only in this example we have based the similarity matrix on the geographical

distance (measured in millimeters with a ruler) for each pair of languages from the center of one to the center of the other. In order to make these geographical distance based similarities directly comparable with the cognate basic vocabulary based similarities, we have transformed the ruler measurements by multiplying each one by 2 and then subtracting the product from 100. This transformation is purely for our benefit since the method works equally well on the untransformed data and produces the same picture in either instance.

Only MIN produces even a trivial structure, and neither MAX nor MIN produces anything even remotely close to the true taxonomic picture. When the data is genuinely not structured taxonomically the method or techniques demonstrates the fact clearly--by falling apart; in computerese the maximum reads G.I.G.O. (garbage in, garbage out)!

III The Salish Results

Our Diameter and Connectedness methods yield the following differences (cf. Figure 4):

1) By the Diameter method, Cm joins Pe, Ka, Sp, Ok before Cr joins those languages; by the Connectedness method, Cr joins them before Cm. Recall that Swadesh kept Cm and Cr separate until the final Interior Division clustering. Dyen places Cm and Cr with Pe, Ka, Sp, and Ok but does not distinguish an order. Jorgensen joins Cm first, then Cr, agreeing with our Diameter method but collapsing the distance between the two nodes to a much smaller distance than the Diameter method.

2) By the Diameter method, Sq joins Nn and Fr after Nt joins Nn and Fr; by the Connectedness method Sq joins Nn and Fr before Nt does.

Swadesh did not include Nt or Sq with Nn and Fr until his Nanaimo Group, Lkungen Group, Sq, and Nt joined in the South Georgia Branch. Dyen joined Sq and Nt into his South Georgia Branch but does not distinguish order of clusterings. Jorgensen agrees with our Diameter method.

3) By the Diameter method, Ti joins the interior languages (cr, Cm, Pe, Ka, Sp, Ok, Sh, Th, and Li) and Be joins the central and north coastal languages (Cl, Lm, Lk, Nt, Nn, Fr, Sq, Pt, St, Cx); by the Connectedness method, Be and Ti join, simultaneously, all the other languages in the final clustering. Recall that Swadesh and Dyen both gave Ti and Be separate divisions. On Jorgensen's diagram, Ti joins the interior languages, agreeing with our Diameter method while Be joins all the other languages in the final clustering.

4) By the Diameter method, the 'South Georgia Branch' (Cl, Lm, Lk, Nt, Fr, Nn, and Sq) joins with the 'North Georgia Branch' (Pt, St, Cx) and Tw; Ni, Sn, and Sk join with the 'Olympic Branch' (Qu, Lo, Sa, Ch, and Cw). Also, the south coast languages (Qu, Lo, Sa, Ch, Cw, Tw, Ni, Sn, Sk) join the interior languages; the central coast languages (Cl, Lm, Lk, Nt, Nn, Fr, Sq) are clustered with the north coastal languages (Pt, St, Cx and Be). Jorgensen has the South Georgia Branch (Cl, Lm, Lk, Nt, Fr, Nn, Sq) with Puget Sound languages (Sk, Sn, Ni) and Tw; Jorgensen joins North Georgia, South Georgia, Puget Sound, and Olympic Branch languages into the Coastal Division at his second to last clustering. Recall that Swadesh kept North and South Georgia branches distinct from each other and from the Olympic Branch while Dyen further differentiated the South Georgia and Olympic Branches; Tw was kept a separate member of the Coast Division by both.

This last Diameter method clustering, representing by far the most significant difference between our results and the three previous classifications, was hinted at by Swadesh in his comments on the earlier discussion of the Salishan languages by Boas (Boas and Haeberlin 1927):

> Boas' division of Salishan into Coast and Interior dialects was surely never intended to be more than a convenient geographic breakdown with only approximate linguistic implications (Swadesh 1950:163).

In spite of this hint by Swadesh, practically all who have since discussed Salish classification have retained the strict Coast, Interior division. On the other hand, we find (by the Diameter method) that at the higher levels of the family-tree, and separating out the effects of borrowing, the more southerly coastal Salish languages form a clustering with the languages of the interior separate from the more northerly coastal languages. The south coast/interior clustering includes Tillamook and the north coast clustering includes Bella Coola.

A brief look at the ethnographic record supports the view that cultural contacts of the sort conducive to linguistic borrowing are inversely related to geographic distance. The ethnographic record also supports the view that cultural contacts between distinct groups are greater where the density of waterways is higher, i.e. in the Puget Sound, Georgia Straits area. Seasonal transhumance associated with seasonally distributed resources, patterns of local group exogamy, trade, and political alliances contributed to frequent contacts among neighboring local groups, especially those interconnected by waterways (cf. Figure 5).

For example, Olsen writes that among the Quinault, central Olympic peninsula, many marriages were intervillage, and because of the samll size

of the Quinault 'tribe' together with kin-group exogamy requirements, intertribal marriages were frequent (Olsen 1936:106). Contacts were with those to the immediate north and south, e.g. the Clallam, Chinnok, and Lower Chehalis as well as with those somewhat further away, e.g. the Tillamook (Olsen 1936:124). Gunther observes that the Clallam only rarely traveled over trails inland; such travel was considered a 'great hardship' (Gunther 1927:212); she observes that the Clallam knew of and had contacts with the Lummi and Swinomish across the straits as well as their immediately adjacent neighbors, e.g. the Makah. On the other hand, the Skagit, Skykomish, and Snuqualmi, who live more inland, across the straits to the east, are 'almost unknown'; Clallam meet frequently with the Snohomish who travel the same waterways. Smith notes that Puyallup-Nisqually had contacts with Snoqualmie, Skagit, Sahaptin speakers, Chehalis, and Twana among others. Contacts were especially along waterways and the more narrow interfleuves. Smith also observes that exogamous marriages were recognized by Puyallup-Nisqually as a means of forming alliances between villages (Smith 1940:42-43).

We have less information on the interior Salish groups but such as we do have indicates, because of the greater distances involved and less dense distribution of waterways, fewer intergroup contacts. For example the Coeur D'Alene were in contact with Spokane, who were near-by, and to a less extent with Pend d'Oreilles and Nez Perce, the latter being Sahaptin [?] speakers. Informants indicated to Teit that Coeur D'Alene married with non-Coeur D'Alene, e.g. the not too distant Columbia, infrequently (Teit 1930:40). Okanagon contacts, including intermarriages, were with those near-by, e.g. the southern Okanagon with Columbia to the south and Shuswap with northern Okanagon.

Bibliography

Dyen, Isidore
1962 The lexicostatistically determined relationship of a language group. IJAL XXVII (3):153-161.

Johnson, S. C.
1967 Hierarchical clustering schemes. Psychometrika, 32:241-253.

Jorgensen, Joseph
1969 Salish Language and Culture.

Sokal, R. R.
1966 Numerical Taxonomy. Scientific American 215(6):106-116.

Sokal, R. R. and P. H. A. Sneath
1963 Principles of Numerical Taxonomy. San Francisco: W. H. Freeman & Co.

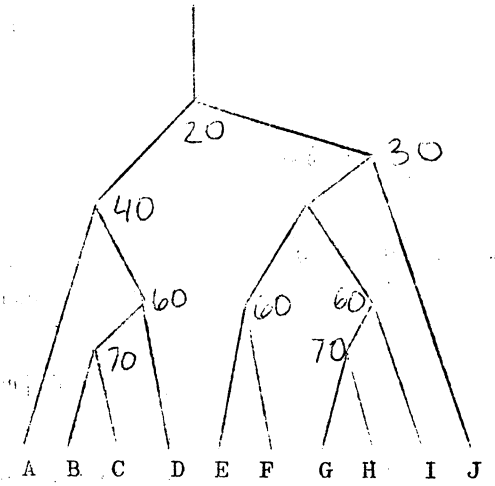
Swadesh, Morris
1950 Salish internal relationships. IJAL 16:157-167.

1951 Diffusional Cumulation and archaic residue as historical explanations. SWJA 7:1-21.

In terms of similarity to the actual relatedness between Salishan languages, I would rank the five family trees given here as follows, from least similarity to highest:

1. Kronenfeld & Thomas, Diameter Method
2. Jorgensen
3. Dyen
4. Swadesh
5. Kronenfeld & Thomas, Connectedness Method

FIGURE E



= Percentages of shared cognates in basic vocabulary.

FIGURE F

Hierarchical Clustering Examples

TRUTH!

The 'real' family tree of a more complicated, but equally non-existant example.

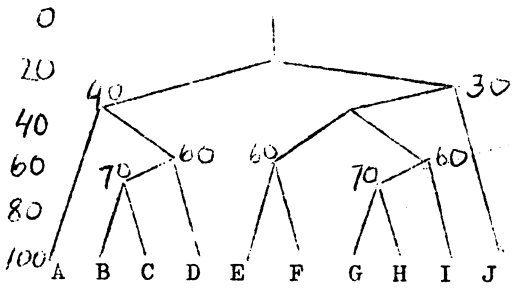
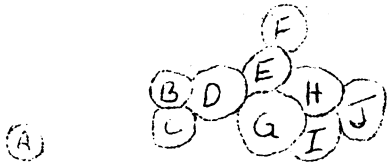


Table of Pairwise Similarities from common Ancestry

	A	B	C	D	E	F	G	H	I	J
A										
B	40									
C	40	70								
D	40	60	60							
E	20	20	20	20						
F	20	20	20	20	60					
G	20	20	20	20	40	40				
H	20	20	20	20	40	40	70			
I	20	20	20	20	40	40	60	60		
J	20	20	20	20	30	30	30	30	30	

Present Geographical Distribution



Affects of Borrowing on Shared Vocabulary

DE +15 EG +15

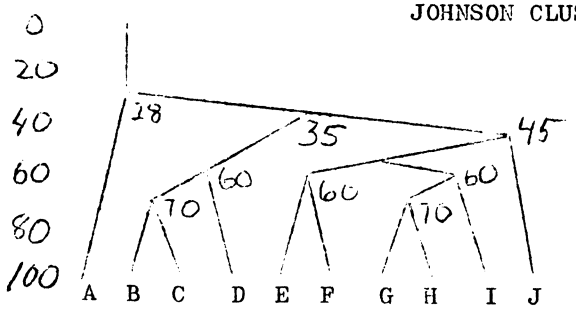
DF +15 A -30% Prorated to all

EF +15

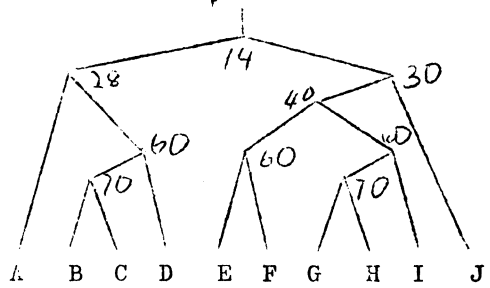
Table of Pairwise Similarities from Common Ancestry and Borrowing

	A	B	C	D	E	F	G	H	I	J
A										
B	28									
C	28	70								
D	28	60	60							
E	14	20	20	35						
F	14	20	20	35	60					
G	14	20	20	20	55	40				
H	14	20	20	20	55	40	70			
I	14	20	20	20	40	40	60	60		
J	14	20	20	20	30	30	30	45	45	

JOHNSON CLUSTERINGS



Minimum Method



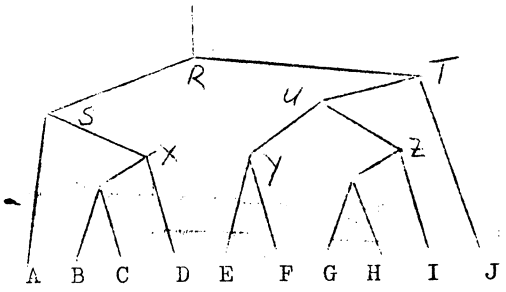
Maximum Method

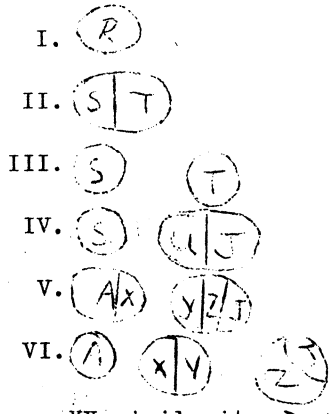
Maximum does better than Minimum by placing A correctly. Otherwise they have essentially the same form as each other--and as TRUTH.

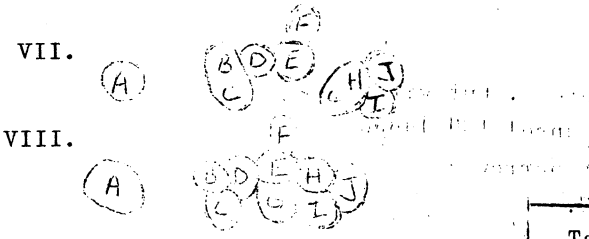
FIGURE G

HIERARCHICAL CLUSTERING EXAMPLES

Use the same "real" family tree and Table of Pairwise similarities from Common Ancestry as in Figure F. Let's repeat the tree (and add labels for the nodes) but not the Table! Now let's give the family a geographic history and let's indicate effects on pairwise similarities via borrowing as we go.







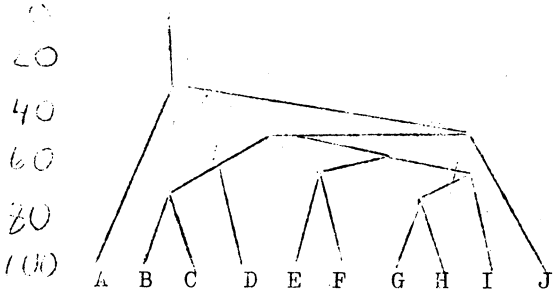
XY similarity → +15
Anticipating that we still have 30% of our diift to go before "the present". Let's prorate figure accordingly
→ +11 = 15-.3(15)

A all -.3 prorated

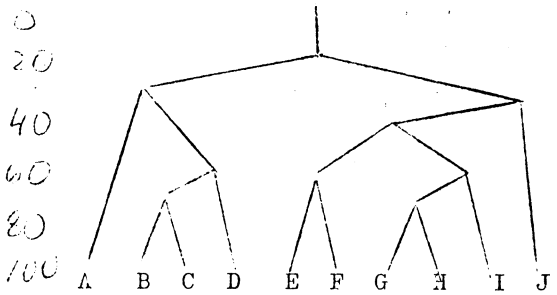
DE +15	EH +15
DG +15	HJ +15
EG +15	IJ +15

Table of Pairwise Similarities from Common Ancestry and Borrowing.

	A	B	C	D	E	F	G	H	I	J	
											A
28											B
28	70										C
28	60	60									D
14	31	31	46								E
14	31	31	31	60							F
14	20	20	20	55	40						G
14	20	20	20	55	40	70					H
14	20	20	20	40	40	60	60				I
14	20	20	20	30	30	30	45	45			J



Minimum Method



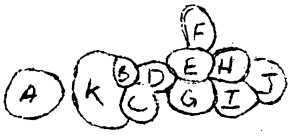
Maximum Method

Notice that the maximum method does better than the minimum method by placing A correctly. And also J! Maximum is better than minimum in two places and essentially matches form of TRUTH. Minimum is nowhere better than maximum.

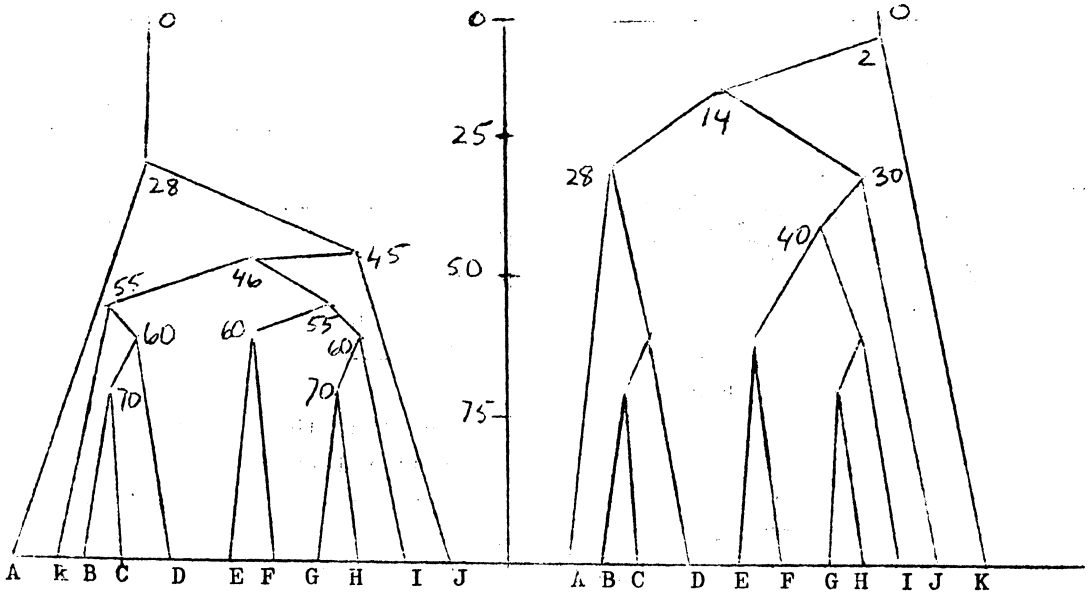
Q.E.D.

FIGURE H

Same as Figure G, but with addition of unrelated language K, which has borrowed heavily with B and C.



	A	B	C	D	E	F	G	H	I	J	K	
A												A
B	28											B
C	28	70										C
D	28	60	60									D
E	14	31	31	46								E
F	14	31	31	31	60							F
G	14	20	20	20	55	40						G
H	14	20	20	20	55	40	70					H
I	14	20	20	20	40	40	60	60				I
J	14	20	20	20	30	30	30	45	45			J
K	5	55	55	2	2	2	2	2	2	2		K



Minimum

Maximum

FIGURE I

cf Figure G

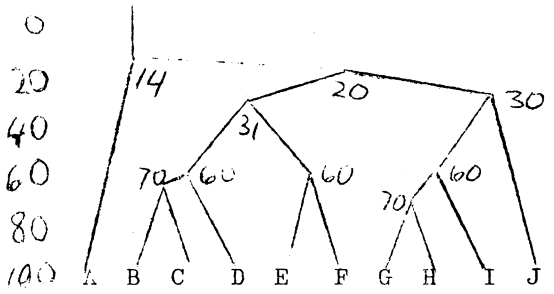
Let us assume that YZ and YJ are lowered .3 by the separation of Y in space as I sort of mistakenly said in Figure G.

Y=E,F	YZ/J=30 (real life)
Z=G,H,I	-.3x: .7x30=.21 from separation
U=YZ	Y/Z=40 (real life)
	-.3x: .7x40=.28 from separation
EG: +15	
EH: +15	.28+.15≒43

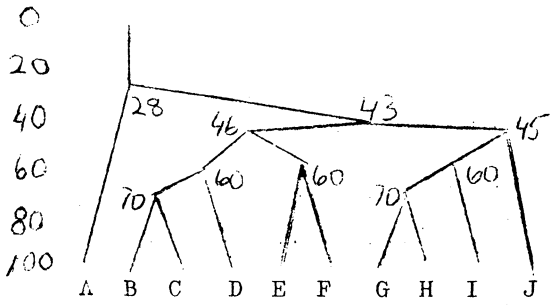
Real tree in Figure F

This table is like the one in Figure G except for (E,F)x(G,H,I,J)

A	B C D	E F	G H I	J	
					A
28					B
28	70				C
28	60 60				D
14	31 31 46				E
14	31 31 21	60			F
14	20 20 20	43 28			G
14	20 20 20	43 28	70		H
14	20 20 20	28 28	60 60		I
14	20 20 20	21 21	30 45 45		J



Maximum Method



Minimum Method

Maximum and minimum are the same. If the historical changes are complex enough and <u>biased</u> enough--the original tree can be irretrievable from the kind of data.

But a) Minimum still not better than Maximum

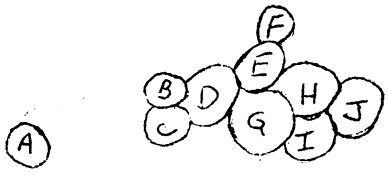
b) these biases are pretty spectacular

FIGURE J

HIERARCHICAL CLUSTERING EXAMPLES

Just to indicate that the preceding results were not merely happenstance, let us take the geographical distribution from Figure F (reproduced here), measure the distances between language centers, construct a table of pairwise similarities from these, and hierarchically cluster the distances.

To make the distance similarities directly comparable with the cognate based basic vocabulary similarities (for <u>our</u> benifit alone, the technique can take either form of data and produce the same result) let us measure the distances in millimeters, multiply the distance by 2, and subtract that number from 100.

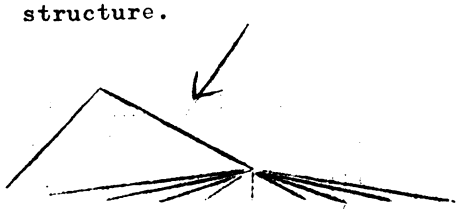


Actual Distances

A	B	C	D	E	F	G	H	I	J	
										A
15										B
17.5	5									C
20	5	5								D
27.5	10	10	5							E
30	15	15	10	5						F
25	10	7.5	7.5	5	12.5					G
32.5	17.5	12.5	12.5	5	10	5				H
32.5	19	15	15	10	15	5	5			I
40	25	20	12.5	12.5	15	10	5	5		J

0
20
40
60
80
100

Only a trivial hierarchical structure.



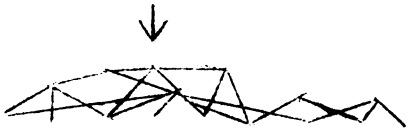
Minimum Method

Transformed Distances

A	B	C	D	E	F	G	H	I	J	
										A
70										B
65	90									C
60	90	90								D
45	80	80	90							E
40	70	70	80	90						F
50	80	85	85	90	75					G
35	65	65	75	90	80	90				H
36	65	70	70	80	70	90	90			I
20	50	60	65	75	90	80	90	90		J

0
20
40
60
80
100

No real hierarchical structure discernable i.e. a mess!



Maximum Method

G. I. G. O. When the data is genuinely not structured taxonomically the method or technique demonstrates the fact clearly--by falling apart!

Q. E. D.

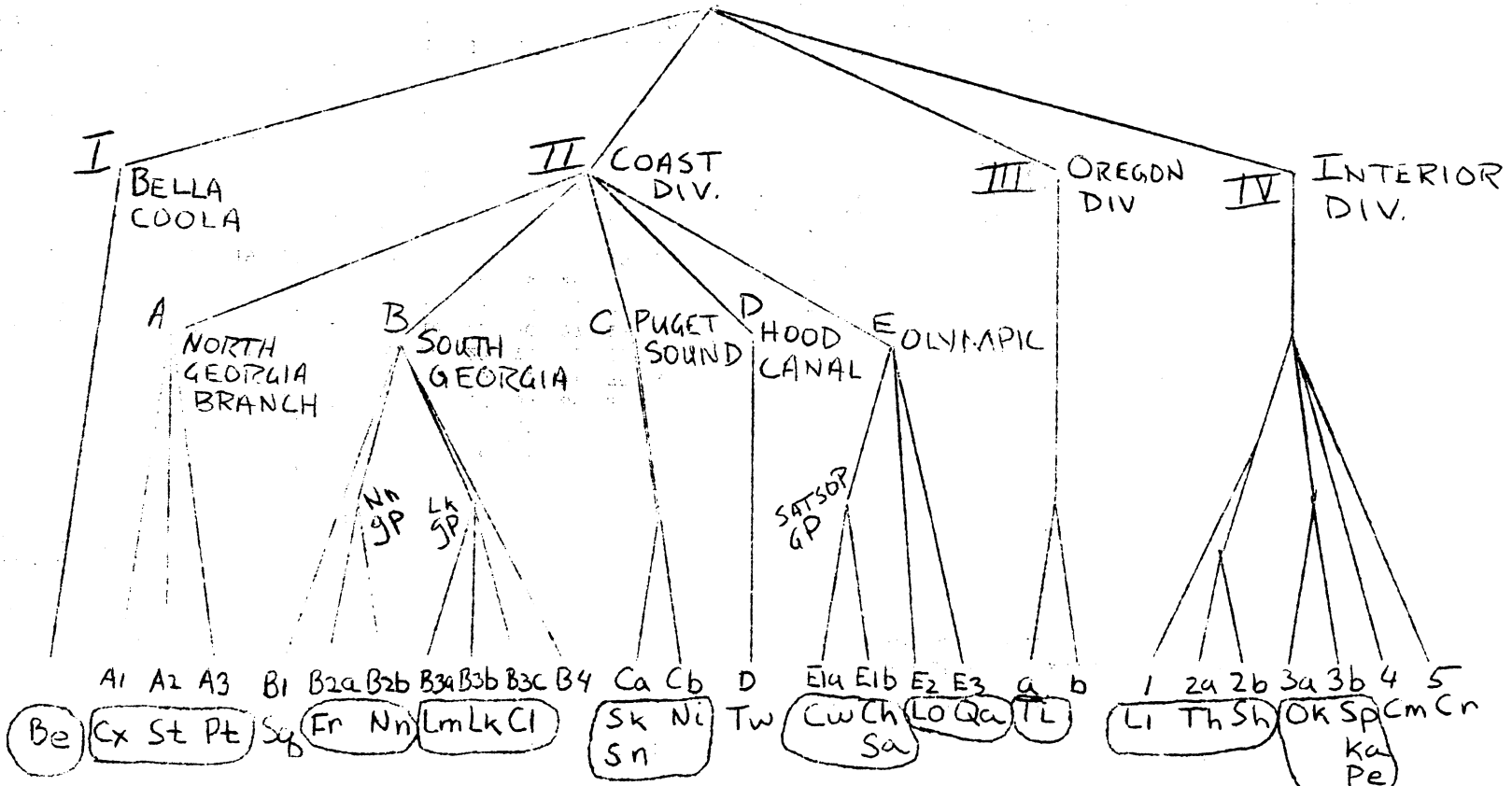
Table 1

Swadesh's Data Matrix

(Dyen 1962:157; cf. Swadesh 1950:159;
cf. Jorgensen 1969:20)

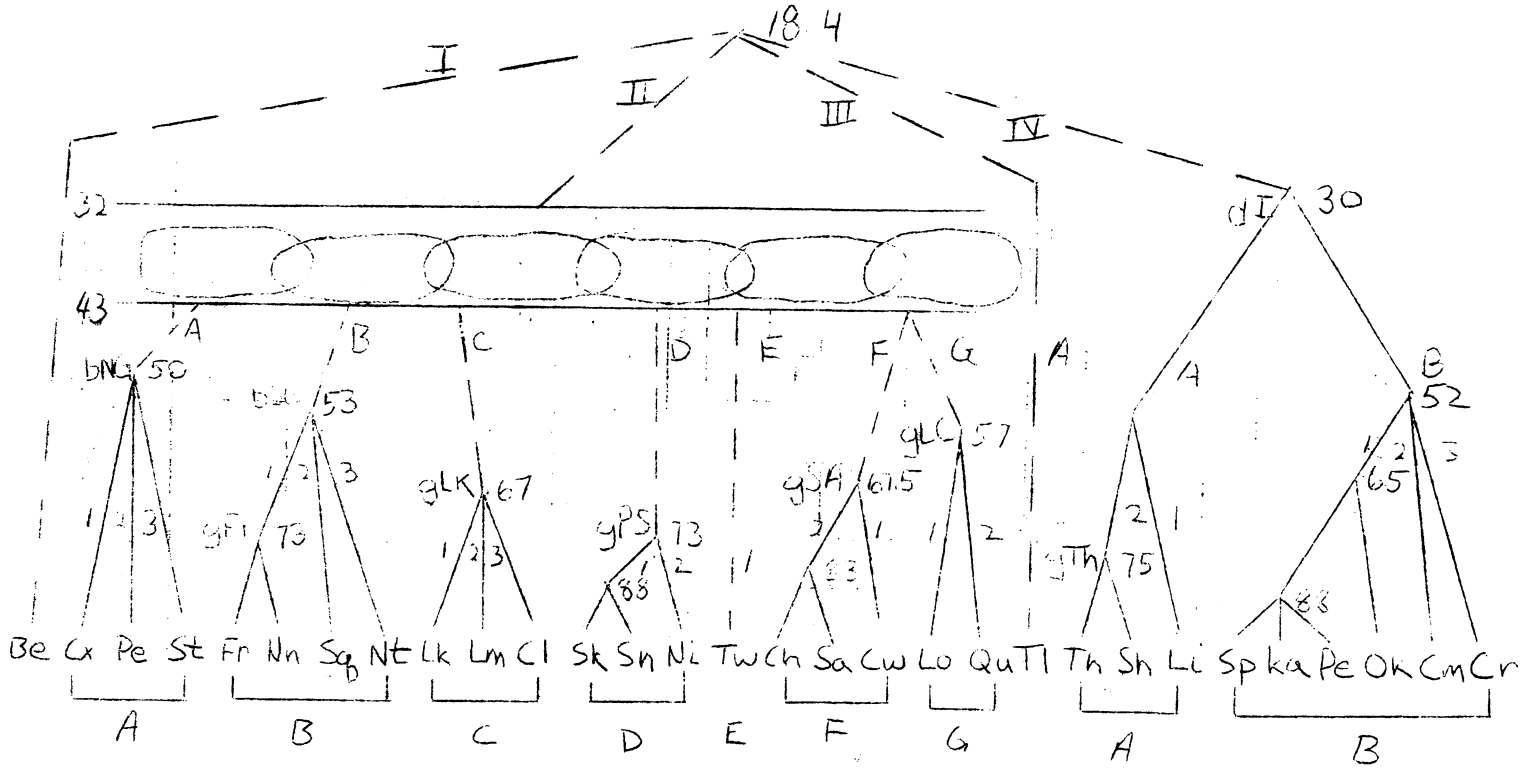
1	Be	Be																											
2	Cx	20	Cx																										
3	St	20	55	St																									
4	Pt	23	45	51	Pt																								
5	Fr	23	35	34	42	Fr																							
6	Nn	22	27	34	38	73	Nn																						
7	Sq	18	27	31	33	50	59	Sq																					
8	Nt	16	31	31	35	58	56	48	Nt																				
9	Lk	23	30	33	34	53	54	38	46	Lk																			
10	Lm	17	28	34	30	45	47	35	47	73	Lm																		
11	Cl	19	27	28	27	36	40	31	40	63	65	Cl																	
12	Sk	12	22	25	21	33	31	30	44	36	41	35	Sk																
13	Sn	16	22	29	23	34	30	29	43	35	37	34	88	Sn															
14	Ni	15	24	25	26	36	33	32	43	34	39	31	71	75	Ni														
15	Tw	15	22	17	19	31	25	25	30	24	27	24	32	33	41	Tw													
16	Cw	15	25	19	18	21	19	15	19	18	21	20	27	28	30	38	Cw												
17	Ch	12	25	20	20	23	21	17	25	19	21	22	32	31	32	39	66	Ch											
18	Sa	14	23	23	20	23	19	17	24	18	20	19	28	28	31	36	59	33	Sa										
19	Lo	13	17	17	14	19	17	16	19	19	20	18	20	20	21	27	43	30	51	Lo									
20	Qu	12	18	16	15	20	20	14	22	21	22	22	24	23	23	22	36	41	43	57	Qu								
21	Ti	14	15	16	15	23	20	20	20	16	16	17	18	20	19	21	16	17	14	15	16	Ti							
22	Li	19	20	22	20	28	25	22	26	20	21	20	21	21	24	23	19	19	21	18	17	19	Li						
23	Th	17	19	24	19	23	22	18	20	20	20	17	21	23	24	21	21	19	18	17	18	22	57	Th					
24	Sh	14	17	20	16	19	17	15	19	19	19	16	19	19	21	20	19	18	18	17	17	18	48	75	Sh				
25	Ok	16	18	20	15	20	18	13	20	20	16	19	19	21	21	22	18	17	17	17	17	22	33	47	50	Ok			
26	Sp	11	15	17	11	16	14	12	18	17	16	16	19	21	23	19	18	15	17	15	13	18	27	36	39	64	Sp		
27	Ka	11	15	17	10	18	15	13	17	17	16	16	20	20	21	20	17	16	15	14	15	17	27	37	40	65	85	Ka	
28	Pe	11	15	17	11	16	14	12	18	18	15	15	20	21	21	20	18	17	17	14	14	18	25	38	39	66	86	93	Pe
29	Cm	12	13	15	10	17	16	14	18	16	18	16	16	19	20	19	20	18	19	17	16	18	25	34	34	54	51	52	52	Cm
30	Cr	12	15	17	10	19	17	14	18	16	17	15	15	17	19	17	18	16	19	17	17	16	23	34	32	44	54	55	55	50	Cr

TABLE 2 SWADESH CLASSIFICATION
(SWADESH 1950: 163-164)



ROMAN NUMERALS AND LETTERS ARE SWADESH'S INDEX

TABLE 3 DYEN CLASSIFICATION (DYEN 1962: 160)

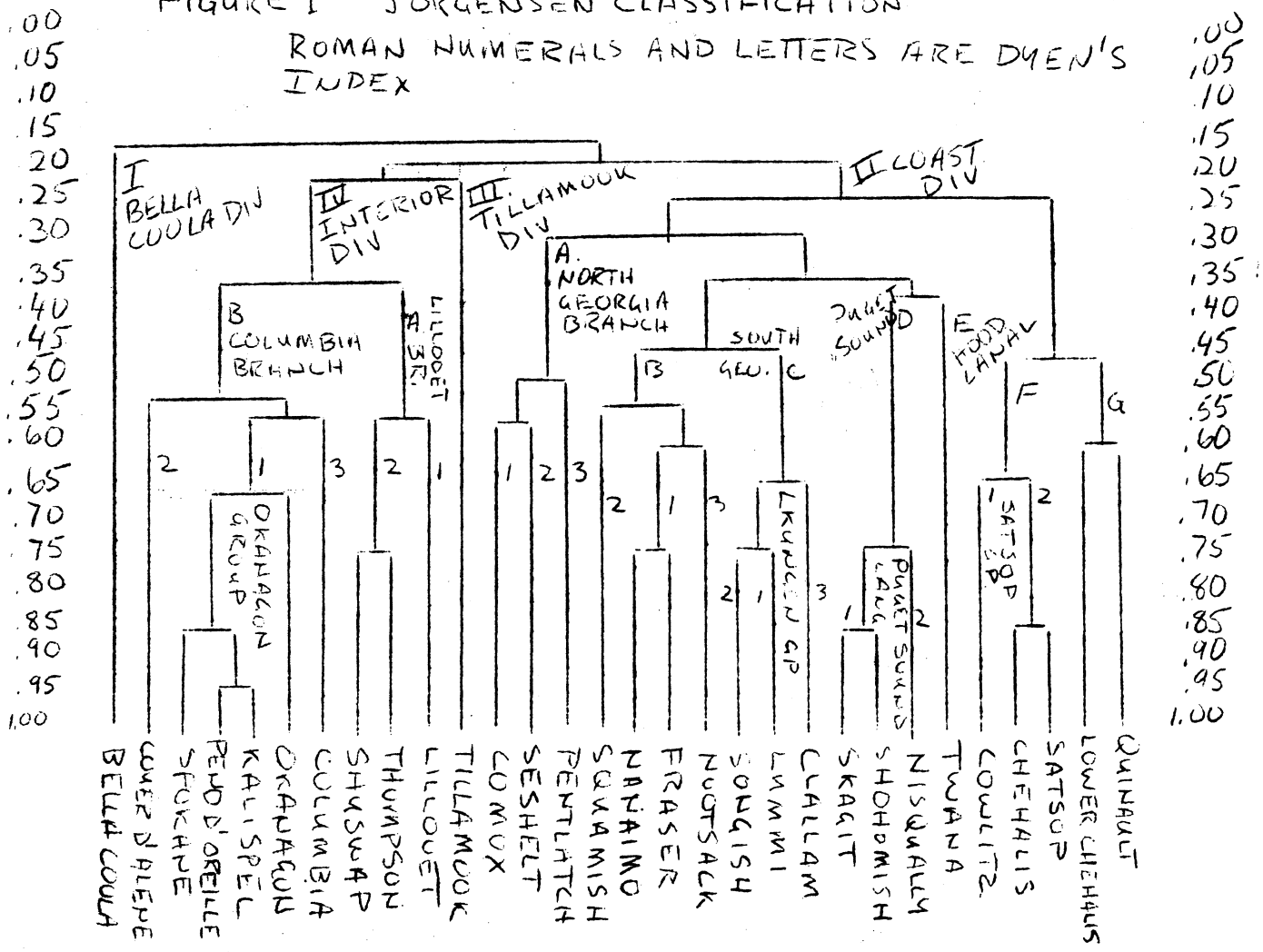


NUMBERS REPRESENT OUR INFERENCE CONCERNING THE PERCENTAGE LEVELS AT WHICH THE INDICATED LANGUAGES CLUSTER

ROMAN NUMERALS AND LETTERS REPRESENT DYEN'S INDEX

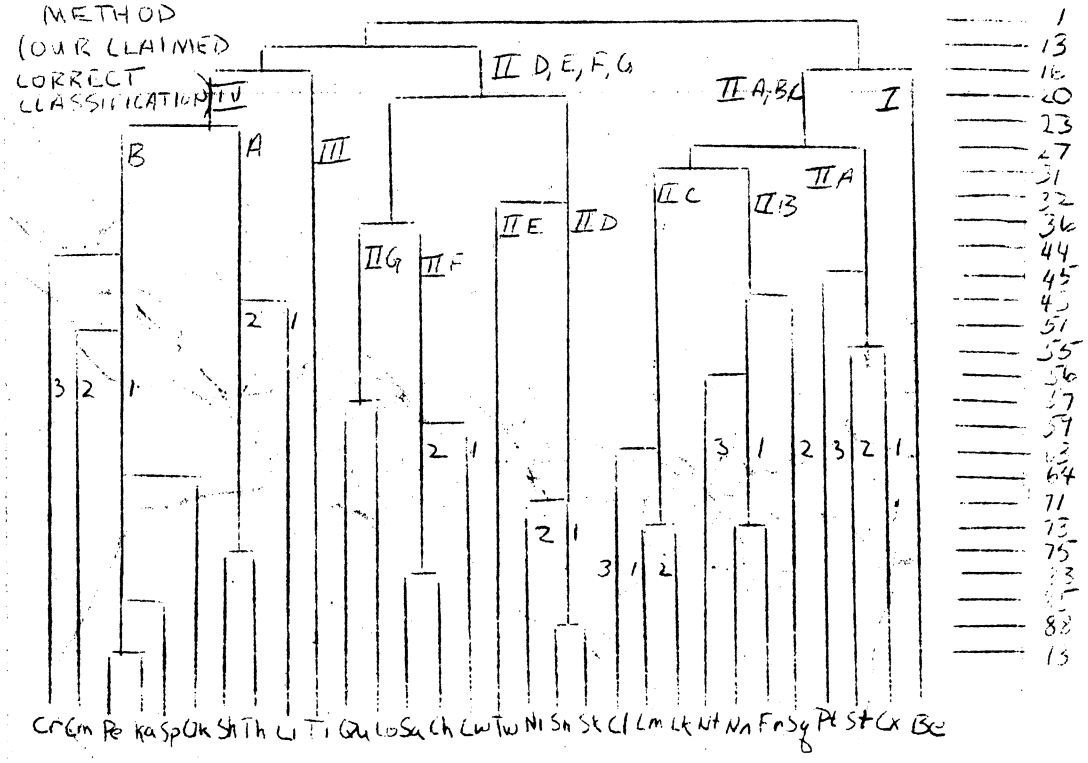
FIGURE 1 JORGENSEN CLASSIFICATION

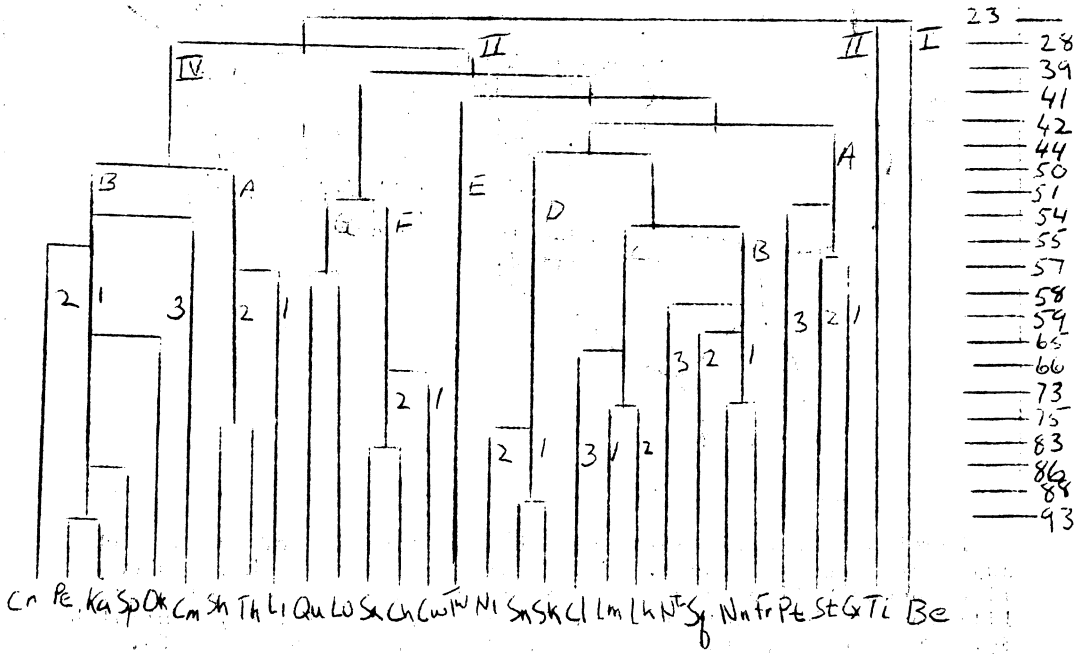
ROMAN NUMERALS AND LETTERS ARE DYEN'S INDEX



OUR CLUSTERINGS
(SWADESH MATRIX)

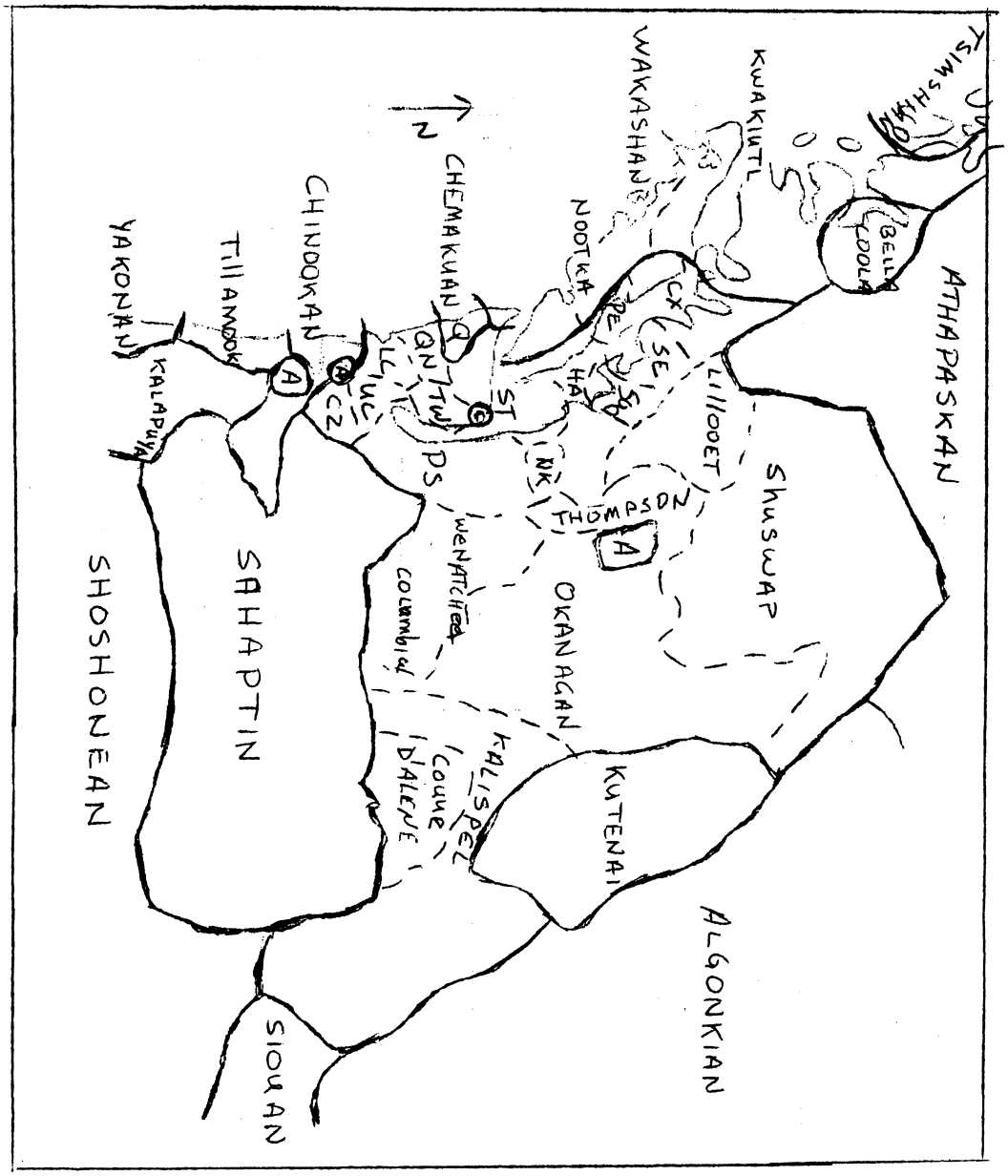
DIAMETER METHOD (OUR CLAIMED CORRECT CLASSIFICATION)





CONNECTEDNESS

CONNECTEDNESS METHOD (SWADESH, cf. DYEN & JORGENSEN CLASSIFICATION)



Portion of Pacific Northwest showing location of Salish and adjacent language stocks. Solid lines enclose stocks, broken lines bound Salish languages (except isolated Bella Coola and Tillamook. Abbreviations: Cx, Comox; Se, Sechelt; Pe, Pentlatch; Sq, Squamish; Ha, Halkomelem; St, Straits; Nk, Nooksack; Ps, Puget Sound; Tw,TWana; Qn, Quinault; LC,Lower Chehalis; UC, Upper Chehalis; Cz, Cowlitz; A, Athapaskan; Q, Quileute; C, Chemakum. Eastern portion of Wenatchee-Columbia boundary may adjoin Kalispel group. Rivers have been omitted.