

## **Internal language classification**

Søren Wichmann

Max Planck Institute for Evolutionary Anthropology & Leiden University

### **1. Introduction<sup>1</sup>**

This chapter will discuss methods of joining languages in groups based on (different degrees of) genealogical relatedness. This criterion is only one among many conceivable ones that may be used to classify languages. Other possible criteria include geography, evidence of language contact or the presence of certain typological features; but these types of criteria will be ignored here. The reason for the limited focus is not only lack of space, but also the special interest that genealogical classification holds within and beyond the language sciences. If languages can be shown to be related genealogically it means that they share a common ancestor. This, in turn, means that something useful may be said about specific human groups in prehistory in some given region through the inspection of the current related languages. But language classification is not only a tool for students of prehistory, it also serves to organize knowledge and direct research. For instance, if it can be shown that a given group of languages are related, then that group of languages may become a target for comparative research. Alternatively, if a given language turns out to lack relatives, then the language in question gains a position of

special interest because of its uniqueness. Thus, language classification is a natural preparatory step before the further in-depth study of languages.

Two aspects of the present contribution will set it somewhat apart from most textbook introductions to the topic. First, language classification will be treated not as *sui generis*, i.e. as a field confined to its own tradition, but rather as a subfield of general phylogenetics, a field which has traditionally been dominated by biology. Therefore, the terminology is often drawn from biology. Secondly, the focus is less on the state of the art and more on potential aspects of the future of the art.

In terms of both goals and methods there are many differences between external and internal language classification. By external classification I refer to the joining of genealogically related languages into maximally inclusive groups. Such maximally inclusive groups are henceforth called families. An example of a family would be Indo-European. Provided that there is sufficient evidence for a higher-level grouping, for instance some version of Nostratic, then this would also be a family, in my use of the term (the example is used for illustrating a terminological issue, and is not meant to imply anything about how I evaluate Nostratic). Germanic, however, never constitutes a family in my use of the term because that this group of languages, as all would agree, is certainly related to some other languages. External language classification has been pursued in many different ways, and a single, consistent method has yet to be applied to all the world's languages. Typically, families have initially been suggested on the basis of certain striking similarities and for some suggestions consensus has eventually been reached that the relationship in question was real, whereas other suggestions have remained controversial to various degrees (see Campbell and Poser 2008: 404-415 for a

comprehensive list of such proposals). The types of similarities have been either lexical or grammatical in nature, if not both, but regardless of the nature of such initial observations consensus concerning the existence of a true phylogeny has never been reached until scholars were able to reconstruct vocabulary and grammar, and to show regular trajectories in the development from a proto-language to its daughters. Such work requires years of dedicated effort applying the comparative method, so there is typically a leap between the initial proposal of a distant genealogical relationship and the acceptance of such a proposal. For instance, it took half a century between the initial proposal of Austroasiatic by Schmidt (1906) until scholars began to establish it more firmly (cf. Pinnow 1959 and papers in Zide 1966), and Sapir's (1913) proposed relationship between Wiyot and Yurok and Algonquian was not generally accepted until the work of Haas (1958).

While long-range comparison clearly merits discussion, little progress has actually been made in this field. Different approaches have been applied, such as the search for shared peculiarities in grammatical organization, which seems to have guided much of Edward Sapir's work, the search for cognates sharing meanings pertaining to a fixed basic vocabulary list (Swadesh 1954), random searches for any possible cognates within a large group of languages (e.g., Greenberg 1987), searches for diagnostic elements, such as similar-shaped pronominals (Nichols 1996) or the comparison of abstract structural features (Dunn et al. 2008). It is not clear which sort of method works best. The only thing which is clear is that each is, at best, only a heuristic. None of them, not even some combination, could deliver the sort of proof for a genealogical relation that would satisfy any historical linguist.

Thus, the establishment of the world's language families has proceeded in a hodgepodge fashion—not by the application of a single heuristic followed by some established probative method. For this reason, and for reasons of space, little more will be said in this chapter about external language classification; the reader is instead referred to the book-length treatment by Campbell and Poser (2008).

Internal classification is the partitioning of a family into smaller units. Any number of terms can be introduced to name groups at different levels of inclusivity, but an analysis of the structures of linguistic phylogenies, to which I shall return below, shows that below the level of maximal inclusivity and above the level of languages there is only one non-arbitrary level of classification, which I refer to as “natural genera”. Once a family is established there are different ways that clades (subgroups) of a family can be established. Clades are groups of languages that are mutually closer related to each other than to languages outside of the group. Two families of methods for establishing clades can be distinguished: character-based and distance-based. A character is a certain phenomenon, such as a cognate word, a phoneme, morpheme, a sound law, an abstract grammatical feature, a syntactic change, etc. which can be present or absent in a given language. Any sort of character may be used to classify languages, but the most widely used within the framework of traditional comparative linguistics are phonological or morphological changes, and within lexicostatistics cognate classes have traditionally been used. Distance-based methods use any sort of measure of distances among languages, establish a distance matrix and derive phylogenies from these. Character- and distance-based methods will be treated in turn in the next two sections.

## 2. Character-based classifications

The framework of the traditional comparative method offers a standard way of partitioning a family into subgroups. The first step consists in distinguishing between plesiomorphies (retentions) and apomorphies (innovations), basing reconstructions on the former and either excluding the latter from consideration when making reconstructions or explaining them as products of changes that eventually derive from phenomena shared with other languages across the family, i.e. underlyingly plesiomorphic phenomena.

The next step of setting up subgroups now consists in looking for synapomorphies (shared innovations), distinguishing them from symplesiomorphies (shared retentions), the latter of which are useless in setting up subgroups. Typically, synapomorphies are chosen from the domains of phonology or morphology since a lot is known about directionality in these domains (in phonology certain changes are known to be more natural than others, such as  $*p > f$  as opposed to the opposite, and in morphology mechanisms such as markedness shift and analogy are well-studied). Campbell (2004: 195) offers examples from sound changes in Mayan languages that have been used for the internal classification of this family.

Often one finds homoplasy, i.e. character states that are independently innovated in two or more groups of languages. This can happen when the innovation in question is a natural one occurring frequently across languages, whether they are related or unrelated; or it may happen because of lateral transfer, i.e. because of borrowing among languages. Homoplasy is the major challenge for internal classification because much is left to the

intuitions of the researcher with regard to determining whether a shared character state can be considered synapomorphic or whether it should rather be interpreted as either an independently occurring natural change or the result of lateral transfer. Lateral transfer, in particular, is a problem for classification because a language change arising in some ancestral language spreads by the same mechanisms as a language change borrowed across groups of languages. Thus, a group of languages comprising the languages A, B, and C, may be defined as a group because of a certain change shared by all three. But it may be difficult, if not impossible, to know whether the change spread among speakers at an early point when A, B, and C constituted a chain of dialects, i.e. at the time of a common ancestor, or whether it spread at a time when the languages were already distinct (Garrett 2006). The best diagnostic for setting up a subgroup is therefore multiple shared innovations: while one change may spread among several languages the likelihood of several such changes having spread at a late stage of complete differentiation is inversely related to the number of changes having occurred. By the same logic, homoplasy is distinguished from synapomorphies: if languages A, B, and C share several changes while languages D, E, F, share several others, and A and F only share one, then it is logical to assume that the change shared by A and F is a homoplasy due to lateral transfer or chance. Geographical data inform such decisions: if a shared character state which is most likely to be due to lateral transfer is found in neighboring languages, then the hypothesis of lateral transfer is strengthened.

The method followed by historical linguists in producing their phylogenies (trees) is in a sense dictated by their model. The model is one of a branching structure where a branch attaches to a root, other branches attach to the first branch, and so forth, and it

incorporates two important assumptions: (1) the assumption of a reconstructible common origin dictates the existence of a root, and (2) the conception of the branching structure itself dictates that there be no lines that connect branches horizontally. While this model has been predominant in historical linguistics ever since it was introduced by Schleicher (1853), it is possible to draw up structure that conform to neither (1) or (2) and yet adequately represent a classification of a set of languages. Figures 1-5 illustrate three alternative classifications of a hypothetical set of four languages. In none of the classifications does a root occur; thus we are dealing with unrooted trees. Since we are not considering the common origin of the four languages we are also not trying to distinguish between synapomorphies and symplesiomorphies. Instead we simply map four different sets of character states depicted as abstract matrices in (1), where the rows correspond to languages and each column is a character which can either be present (1) or absent (0) in a given language.

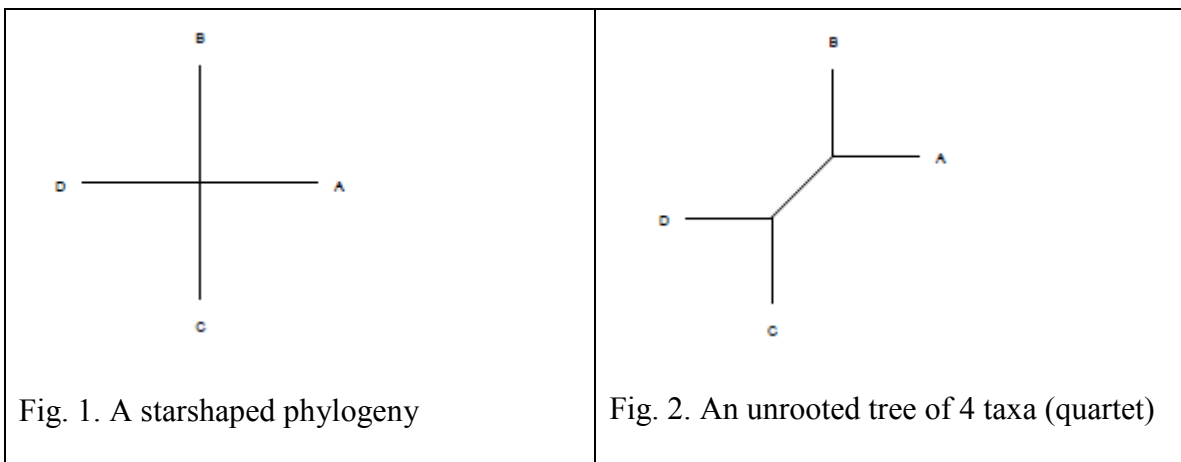
(1) Matrices defining different relationships among four hypothetical languages

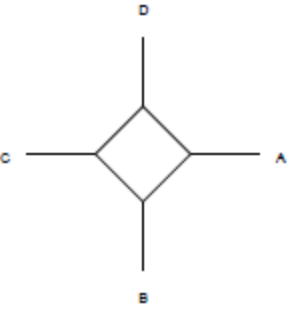
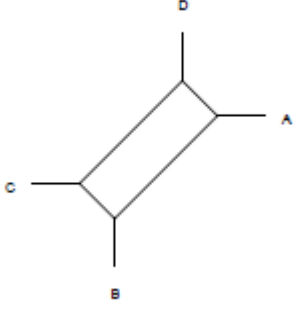
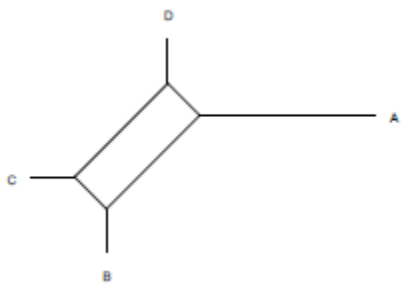
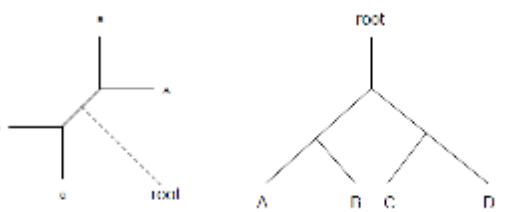
I	II	III	IV	V
A 1000	A 10000	A 100001	A 10000111	A 10000111111
B 0100	B 01000	B 010000	B 01000000	B 01000000000
C 0010	C 00101	C 001010	C 00101000	C 00101000000
D 0001	D 00011	D 000111	D 00011111	D 00011111000

In matrix I each language has its own unique characteristic not shared with one of the other languages. There is therefore no internal structure to the tree—it is completely star-shaped (fig. 1). In matrix II a character has been added which is present in C and D but absent from the two other languages. Now the tree gains some structure: C-D group together against A-B (fig. 2). In III yet another character has been added which is shared by A and D to the exclusion of B and C. In the kinds of trees that linguists traditionally operate with a conflict arises which cannot be solved because the data point in different directions. Are we to join A-B against C-D because of the fifth character or are we to join A-D against B-C because of the sixth character? In the algorithm called Split Decomposition, which is implemented in the phylogenetic software SplitsTree (Huson and Bryant 2006), no attempt is made to somehow resolve the conflict. Instead a square is inserted from whose corners branches lead to each of the four languages (fig. 3). This graphically depicts reticulations in the tree; the structure is less treelike the more such reticulations are found. A resolution of the conflict may be obtained by collapsing parallel edges of the box. Since there is just one box with two sets of parallel edges, two possible resolutions are possible, and since the edges are equally long, one solution is as plausible as the other. This situation changes for matrix IV. Here two extra characters have been added which unite A-D against B-C such that there are now three characters supporting this configuration and only one character supporting A-B against C-D. This leads to the network in fig. 4, where the sides of the box are not equal anymore. Collapsing the longer two edges amounts to ignoring the single character that supported A-B against C-D. In practice, this is what historical linguists often do when they decide that a shared phenomenon which is distributed in an unusual way is most likely

homoplastic (due to borrowing or chance). They may be well advised in doing so, yet the resulting clean tree structure represents a loss of information since, after all, homoplasies are also of interest to the student of language history. In the last matrix (V) yet another set of characters has been added which serve to conclude this brief introduction to phylogenetic structures. This is a set of three characters uniquely present in language A. What these produce is a lengthening of the branch leading to language A (fig. 5). In a traditional linguistic family tree branch lengths are not distinctive: the branches of a tree are simply arranged in whatever way is graphically most convenient. But by using precise algorithms that turn data into trees, however, it is possible to depict the distinctiveness of each node, including terminal nodes such as the one leading to A in fig. 5. In a tree or network that has distinctive branch lengths it is possible to read off information about the amount of evidence that sets off a node defining a subgroup or a single language from the rest of the languages. In contrast, in a tree or network that only depicts a topology, i.e. a mere arrangement of nodes, this kind of information is lost.

Figures 1-6. Different relations among 4 hypothetical languages



 <p>Fig. 3. A network of four taxa</p>	 <p>Fig. 4. Another network of four taxa</p>
 <p>Fig. 5. A network illustration distinctive branch lengths</p>	 <p>Fig. 6a-b. Two versions of the same rooted tree</p>

The hallmark of the comparative method is reconstructions of ancestral states. Since the method operates with a hypothetical proto-language this reconstructed entity carries over to the trees we are used to seeing. If we return to the situation depicted in Figure 2 where a tree is partitioned in two groups we can imagine a proto-language that accounts for the commonalities between the ancestors of respectively A-B and C-D, and this then gets inserted as the root intermediate between A-B and C-D, as depicted in Figure 6a. This tree is equivalent to the more typical graphic depiction in Figure 6b.

In contrast to the reticulated networks of Figures 3-5, Figures 2 and 6 represent perfect phylogenetic networks, i.e., trees based on evidence concerning character states that are not in contradiction with one another phylogenetically speaking. In a non-rigorous application of the comparative method such trees are often set up in despite the knowledge that there are facts contradicting the treelike structure. A truly perfect phylogenetic network, in contrast, would require a rigorous selection of characters whose changes down the tree are not repeated on different nodes. Such a selection of characters would therefore represent explicit arguments for the particular phylogenetic structure claimed to characterize a given language family. For Indo-European, Nakhleh et al. (2005a,b) have presented such a selection, producing a perfect phylogenetic network for this family. Since the characters in question are the kinds of phonological and morphological innovations identified throughout the history of Indo-European comparative linguistics there is nothing new in their contribution, except that it sets more rigorous standards for the passage from data to inferring a tree. Given the advanced nature of Indo-European studies and the combination of a well-tested method in historical linguistics with a modern, rigorous phylogenetic approach this work sets an example for scholars working on other families. Unfortunately, with a few exceptions, other language families have not been studied in the sort of depth where so many details about phonological and morphological developments are known as is the case for Indo-European. Therefore many classifications take recourse to lexicostatistics, which serves as a shortcut in language classification inasmuch as it draws upon a highly selective dataset.

Different lexicostatistical methods have developed, but they share the same sort of dataset, which is a set of words sharing the same meanings across the languages compared. Typically the standard 100-item Swadesh list or some variant thereof is used. Traditionally lexicostatistics has been distance-based, using percentages of shared cognates for each language pair. But the data are discrete characters. It is generally the case that a matrix of characters can be transformed to a distance matrix, but since such a transformation typically represents a loss of information it should probably be avoided, if possible. To illustrate how a character-based approach to lexicostatistics works let's consider a set of four Swadesh list items for four Germanic languages.

(2) Four Swadesh items for three Germanic languages

	Danish	Swedish	Dutch	English
'person'	menneske	människa	mens	person
'skin'	skind	skinn	huid	skin
'fire'	ild	eld	vuur	fire
'leaf'	blad	löv	blad	leaf

A linguist without expertise in the Germanic languages would derive the matrix of cognate classes in (3) from the data in (2):

(3) Cognate classes corresponding to (2)

	Danish	Swedish	Dutch	English
'person'	a	a	a	b

‘skin’	a	a	b	a
‘fire’	a	a	b	b
‘leaf’	a	b	a	b

Although Swadesh recommended not drawing upon knowledge about loanwords for one’s decision it is clear that if we do, much noise in the data can be avoided. This is a way to avoid a minor evil of skewing the relation among related languages such that those which have had more contact are joined closer to another because of loanwords; more importantly, perhaps, a major evil of joining unrelated languages more closely because they happen to both have borrowed basic vocabulary from some major languages such as Arabic (true of many languages in Eurasia and Africa) or Spanish (true of many languages in Latin America) can be avoided. As regards the examples in (2), English has borrowed *person* from Middle French and *skin* from Old Norse. The corresponding forms may profitably be treated as if English lacked words for ‘person’ and ‘skin’.

Some phylogenetic algorithms take discrete characters as input and may be applied to derive trees from abstractly encoded cognate classes. Often there is a limitation on the number of different character states allowed for in the input. This turns out to be a problem for larger families where there may be dozens of different etyma for a single basic vocabulary meaning. This problem is solved by recoding each character as a number of binary characters corresponding to each character state. To use the Germanic example for an illustration of this procedure, the character state represented by the cognates *blad* and *blad* in Danish and Dutch is now treated as one whole character, where Danish and Dutch score 1 for “present”, while English and Swedish score 0 for “absent”.

Similarly, the character state represented by Swedish *löv* and English *leaf* is recoded as a character, where Swedish and English score 1, while the two other languages score 0.

This produces a larger matrix, as illustrated in (4).

(4) Character states of (3) recoded as binary characters

	Danish	Swedish	Dutch	English
‘person-1’	1	1	1	0
‘person-2’	0	0	0	1
‘skin-1’	1	1	0	1
‘skin-2’	0	0	1	0
‘fire-1’	1	1	0	1
‘fire-2’	0	0	1	0
‘leaf-1’	1	0	1	0
‘leaf-2’	0	1	0	1

A variety of phylogenetic algorithms and implementations thereof (typically in software freely distributed on the Internet) are available for turning such matrices into phylogenies. Among the currently most sophisticated and apparently most adequate is so-called Bayesian inference (Huelsenbeck and Ronquist 2001), which is a complicated algorithm that generates different trees and selects the set of most adequate ones by measuring their “posterior probabilities”. Other sorts of algorithms tend to work the other way around, i.e. by starting from the data and subsequently fitting trees to them rather than starting from trees and, working through different trees, finding one or more that

have a maximal likelihood given the data. Computational phylogenetics is a rich and rapidly developing field. For a general in-depth introduction see Felsenstein (2004). Introductions to linguistic computational phylogenetics may be found in Nichols and Warnow (2008), Wichmann and Saunders (2007), and McMahon and McMahon (2005).

### **3. Distance-based classifications**

From its outset, lexicostatistics has operated with distances among languages as a criterion for their classification. When Swadesh (1950) introduced the method, he measured cognate percentages on a standard wordlist for Salishan languages. The kind of representation he chose for the results then as well as in subsequent works was a rather inelegant format, where language names were put in boxes whose mutual arrangement was intended to indicate their genealogical relations. To facilitate the task of arranging the boxes he arranged (or binned, using the technical term) the lexical distances in discrete groups from zero to some maximum. Language groups separated by two units of time depth were put in adjoining boxes with a common boundary, separations of three units were shown by a narrow space between the boxes, and separations of more than three units by a wide space. The procedure constituted a primitive sort of phylogenetic algorithm. Had Swadesh attempted to draw up tree structures more similar to those standardly used, his method would have looked less alien in the eyes of the historical linguistics community and it would have been easier to compare his results to those of other historical linguists. Unfortunately, however, the development of methods to create

phylogenetic trees from distance data was still in its infancy around the time of Swadesh's untimely death in 1967. An early algorithm which is conceptually so simple that it can be applied by hand is UPGMA or Unweighted Pair Group Method with Arithmetic means (Sokal and Michener 1958). The first step in the method consists in joining the pair of taxa, A and B, that have the smallest distance and then redefining this pair as a taxon in itself. Then the distance matrix is recalculated by setting the distance from the new A-B taxon to each other taxon equal to average of the distance from A to the other taxon and from B to the other taxon. Now the joining of closest taxa is repeated, and the procedure continues until all taxa are joined in a tree. While simple, this algorithm has the disadvantage that it assumes that rates of change are equal. Among many other distance-based algorithms that do not make this assumption the one called neighbor-joining (Saitou and Nei 1987) is currently the most widely used.

Counting cognates on a standard meaning list is mostly straightforward for a relatively young family, but it becomes tenuous for distantly related languages, where it is entirely left to the linguist to decide, based on acquired knowledge and intuitions about typical sound shifts, what constitutes and what does not constitute a possible cognate. Moreover, a linguist comparing wordlists from languages not normally assumed to be related would suspend normal evaluations of cognacy in light of the knowledge that the languages compared are not assumed to be related. In this case it would be even more difficult to remain objective. To overcome the subjectivity involved in cognate identification different computational approaches have been developed (Oswalt 1970, Guy 1980, Goh 2000, Kondrak and Sherif 2006, Brown et al. 2008). Such methods, however, have so far not had any practical application in the classification of languages.

More recently, another approach to the computational classification of languages based on lexical information has developed. The approach is based on measurements of phonological distances among words, and pays no attention to whether they are cognate or not. While there are different ways of measuring such distances in the literature, they all take as their point of departure the Levenshtein or “edit” distance, which is defined as the minimal number of substitutions, deletions, and insertions which it takes to get from one word to the other (Levenshtein 1966). While initially applied to dialectological data (Kessler 1995, Nerbonne et al. 1999), Serva and Petroni (2008) and Holman et al. (2008a) have used Levenshtein distances to classify languages. The advantages of the method are that it is computationally much less costly and conceptually simpler than cognate identification procedures. It therefore holds promise to become an effective tool for producing provisional classifications of languages and dialects. In fact, using the subset of the 40 most stable items on the 100-item Swadesh list which was identified by Holman et al. (2008b), Müller et al. (2009) have succeeded in producing a tree based on lexical distances among 3384 languages and dialects in the world and are continuously updating their results as the database of the so-called Automated Similarity Judgment Program (ASJP)<sup>2</sup> expands. Wichmann et al. (2009a) provide some statistics on the comparison of the ASJP classifications with those of experts, showing that the agreement is quite variable, but that there is a tendency for the amount of agreement to be inversely related to the size of families, suggestion that for large families that are not yet well worked out in terms of their historical configurations, the new method may be of utility as an approximation to the kinds of results that might eventually be reached with more in-depth work within the framework of traditional comparative linguistics.

The ASJP method also allows for setting up objective, arbitrary criteria for different levels of genealogical groups. For instance, an IE-level group could be defined as a group of languages having a maximal time depth corresponding to that of Indo-European. Such an approach is likely to yield language groups that are either uncontroversial or ought to be uncontroversial. In addition, many other types of clustering of the world's language families would be possible, including one which I discuss in the next section.

#### **4. Subgrouping for comparative purposes**

This section will explore how subgroups of languages can meaningfully be established such that they are comparable across families. Two different strategies will be considered, where the first is a strategy to establish groups that are comparable in age across families, the age being chosen arbitrarily, and the second is a strategy to find an intermediate level across families where a partitioning emerges naturally rather than being arbitrarily posited.

The first of these two strategies has been applied in work by Matthew Dryer. In order to establish genealogically balanced language samples for typological purposes Dryer (1989: 267) introduced the notion of “genera”, which was defined as “genetic groups roughly comparable to the subfamilies of Indo-European, like Germanic and Romance.” In some cases a genus is also a family. In Dryer (2005: 584-644) more criteria were included in the definition. Here it is said that “a genus is a group of languages whose relatedness is fairly obvious without systematic comparative analysis”; “a genus in one

family is intended to be comparable in time depth to genera in other parts of the world”; and “if there is evidence of time depth of groups, the genus would not have a time depth greater than 3500 or 4000 years”; finally, Celtic is given as the prototype for a genus. A specific age for Celtic is not offered, but it follows from the discussion that its age is considered to be close to the upper bound for genera. While Dryer admits that his list of genera is really only based on educated guesses, it is possible to test the relative time depths of his genera using the ASJP data mentioned in the previous section. Currently the database allows for assigning relative ages to 278 of Dryer’s genera. These ages are found by partitioning the given group of languages in two using the structure of a neighbor-joining tree rooted by its midpoint, where the midpoint is defined as the point in the structure equidistant between the two most divergent members. The average lexical distance is then found for all language pairs whose members are separated by the root, and this average distance represents the relative age (an absolute age may also be calculated given a set of calibration points where a linguistic splitting event is associated with a known date, but this is a problematical area of research which I shall not bother to enter). The result is that 198 of Dryer’s genera have ages that are lower than that of Celtic (as estimated using data from currently spoken or recently extinct languages), while 79 have ages that are higher. While many of the latter are perhaps still within reasonable bounds of Dryer’s definition, they include many which are actually older than Indo-European. Some of the dates are doubtful because trees are skewed such that one of the major branches contain just one language, which raises the influence of this single language on the date out of proportions and increases the margin of error. In other cases problems relating to the data, such as complex morphologies that have not been taken

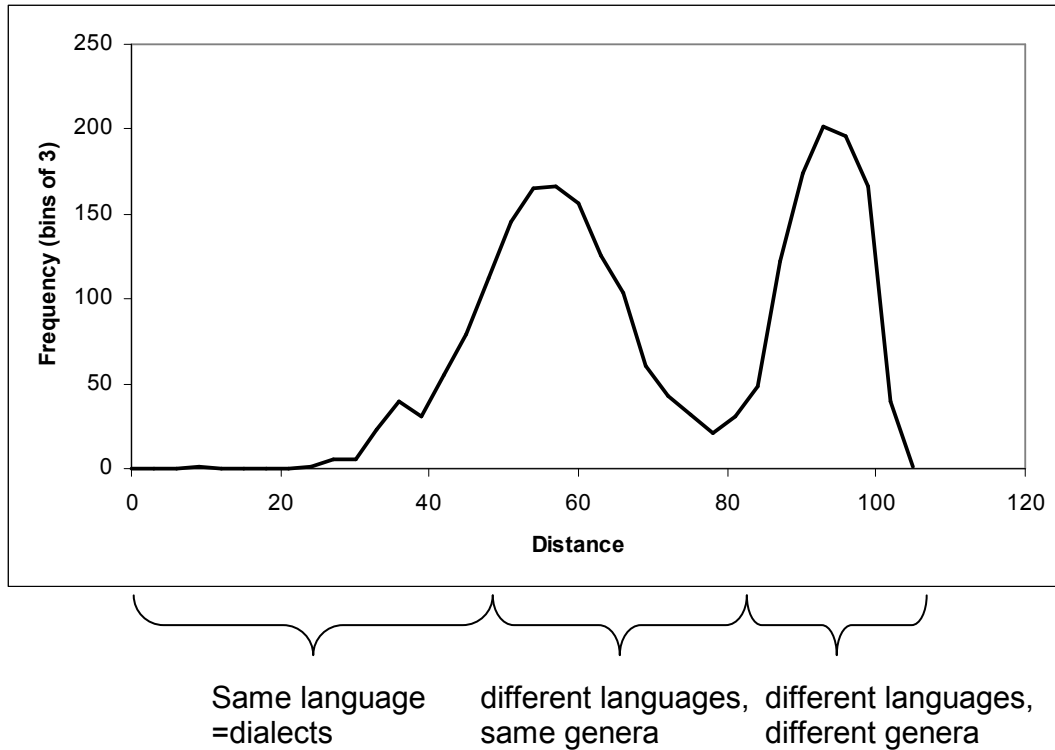
properly into account, may have inflated the age estimate. But in the following cases age estimates higher than Indo-European are well supported—in Africa: the West Chadic subgroup of Afro-Asiatic, the Ubangi, Southern Atlantic, Northern Atlantic, Kwa, Gur, and Adamawa subgroups of Niger-Congo; in the Papuan realm: East Geelvink Bay, Morehead and Upper Maro Rivers, the Wapei-Palei subgroup of Torricelli, the Madang, Eastern Highlands, and Dani subgroups of Trans-New Guinea; and in South America: the Ge-Kaingang subgroup of Macro-Ge. Other families are over-differentiated. Sometimes this is because a subgroup which would be too young to count as a genus is seen by experts as directly branching from the root of the family tree and therefore, by being excluded as a member of some other genus, must by necessity count as a genus in itself. Such cases are inevitable, but there are more problematical cases, where a family is divided into genera even though the family itself is younger than Celtic. These include Dravidian, Tai-Kadai, Chukotko-Kamchatkan, and Wakashan. Thus, a thorough revision would be needed to produce an adequate list of genera in the sense of Dryer (1989, 2005).

The previous paragraphs treated the issue of establishing genera across families based on an arbitrary age criterion. I now turn to the issue of whether there is support from the internal configurations of language families for different levels of classification, and I shall argue that there is support for a notion of natural genera. Unlike a Dryer-type genus, a natural genus is not defined arbitrarily by an age criterion, but is found individually for a given language family by a novel method presented in Wichmann et al. (2009b). Such natural genera have varying ages across families, but they strongly tend to emerge around the time when a proto-language has fragmented into different languages which are beginning to form their own dialects, i.e. when each daughter of the proto-language is

beginning to form its own lineage. The finding that natural genera emerge from the immediate daughters of a proto-language is intuitively appealing, as is the finding that it is possible to distinguish languages from dialects, but, crucially, there is nothing subjective about the method which I propose to identify such intermediate levels of classification.

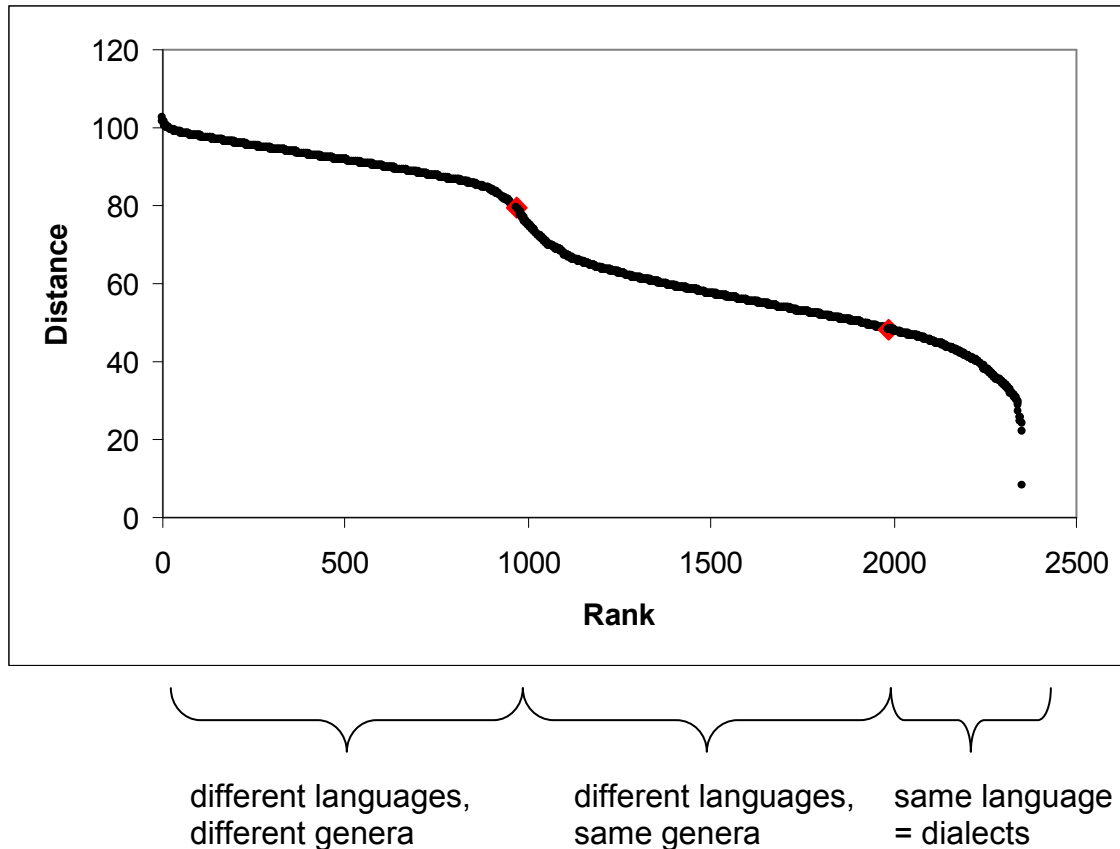
If there are transitional levels in phylogenies corresponding to something like genera it should be possible to find them by plotting distances<sup>3</sup> among all language pairs in the family. Language pairs whose closest ancestor is the proto-language itself should have distances normally distributed around some mean, since they would have the same age of separation. The distances for language pairs whose closest ancestor is somewhere further down the tree would not fit this distribution and a so-called skew normal distribution would arise. Figure 7 shows what such a distribution looks like for Uto-Aztecan. It can be appreciated that as one moves from left to right until coming close to an 80% distance, the distribution begins to no longer be normal (symmetrical around a mean). Moving further to the left there is another peak. This would correspond to different languages within genera. Again, in the left part of the curve, there is transition, this time presumably from languages to dialects.

Fig. 7. A frequency plot of distances for pairs of Uto-Aztecan languages



The exact positions of the transitions are hard to discern from a distance-by-frequency plot such as that of Figure 7. For this purpose another type of plot, shown in Figure 8, is more useful. It is based on the same data but represents the transitions among the different regimes in an alternative, perhaps more vivid way.

Fig. 8. A rank-by-distance plot for Uto-Aztecan



The way in which the transitional points in the curve of Figure 8 are found can be pictured as a problem of fitting the largest possible box under each of the relevant segments of the curve. Going from left to right we see the first transition taking place somewhere just before rank 1000 is reached. An exact point can now be found by finding the maximum of the product of distance and rank for the segment where the rank is lower than 1000. The distance corresponding to this maximum turns out to be 79.4%. The next transitional level is found in a similar way for the segment where the rank is higher than 1000, and turns out to correspond to a distance of 48.4%.

In Wichmann et al. (2009b) 18 plots were produced for families that are sufficiently well attested in the ASJP database to be amenable to this kind of treatment. The relative

ages of the families were determined from distance measures as described towards the beginning of this section. Then the relative ages corresponding to the points of transition between languages in different and same genera were subtracted, and the average time from the proto-language to the emergence of genera could be determined. This average age was a little less than the relative age our method assigned to Slavic. Translating it into an absolute age would require calibrations that are bound to be controversial, but, as a matter of fact, different calibration do not give widely different results—all point to a lifetime of protolanguages of somewhere around a millennium, perhaps a bit more. Given the clear existence in all the world's larger families of transitional points for genera such as the one shown in Figures 7-8 for Uto-Aztecan, it may be inferred that there is such a thing as natural genera, and given their average age, it may be inferred that genera arise at about the time when the immediate daughters of the proto-language begin to form their own offspring. Since there are no other transitions until the family-to-dialect transition is it only the highest splits in the trees that correspond to genera.

The ASJP dataset currently only contains a few families with a large representation of dialects; the Uto-Aztecan dataset plotted in figures 7-8 is unique in that the majority of the speech varieties are very close (nearly all of them being varieties of Nahuatl), while clearly distinct languages form a minority. Thus, presently not much may be inferred about a typical lexical distance or age separating languages and dialects. But in a near future, using this methodology, it should be possible to establish that there is, in fact, a meaningful distinction to be made between languages and dialects, and then to define this distinction quantitatively.

## 5. Outlook

External language classification was treated only cursorily in this chapter. History has shown that there is often a great leap from the initial proposal of a family relation to the point where the relationship has become accepted and generates a field of scholarship. There is currently an abundance of interesting proposals concerning genealogical relations which wait to be fleshed out by more evidence. Until then, such proposals are bound to be controversial. There is clearly progress ahead in this area, but it looks to be as slow as it has always been, since historical linguists, while they have developed several interesting heuristics over the past century, have failed to produce methods that would rapidly prove a distant relationship to the satisfaction of the entire community of historical linguists.

Where linguists tend to agree more is with regard to the internal classification of language families. There are discussions over this as well, but controversies tend to be more controlled because there are clearer criteria for internal than for external classification. This relative methodological success opens up an area of study which has so far largely been neglected within historical linguistics, namely the study of family trees within the wider framework of phylogenetics. Given a rich set of study objects, namely all the phylogenies for the world's language families, we may begin to discern shared or distinctive structural patterns. For instance, we may wonder about whether such trees are more or less balanced in comparison to, say, biological trees (Holman 2009), whether they exhibit natural clusters revealing something about the population dynamics

that produced them (Wichmann et al. 2009b), whether they show effects of increased rates of change as populations diverge (Atkinson et al. 2008), and so on. To address or even ask such questions requires quantitative thinking and ways of transforming language data into numbers. This brings lexicostatistics, which is often seen as nothing but an inferior approach to language classification, into a renewed focus, because what this method does is precisely to transform language data into numbers. Nothing, however, precludes us from developing other quantitative approaches to language comparison, and the field is certain to see interesting developments in this direction in the future.

## Notes

1. My sincere thanks go to Johanna Nichols and Eric W. Holman for helpful comments on this paper.
2. I am grateful for my fellow members of the ASJP consortium, Dik Bakker, David Beck, Oleg Belyaev, Cecil H. Brown, Pamela Brown, Matthew Dryer, Dmitry Egorov, Pattie Epps, Anthony Grant, Eric W. Holman, Hagen Jung, Johann-Mattis List, Robert Mailhammer, André Müller, Uri Tadmor, Matthias Urban, and Viveka Velupillai, for permission to use the database contents and software developed by these scholars in some of my analyses.
3. The distance measures used for the plots in Figures 7-8 are based on Levenshtein distances, but are modified to take into account variable word lengths and accidental phonological similarities. The exact nature of these modifications needs not concern us

here (see Bakker et al. 2009: 171 for a full description), but to avoid confusing the reader it needs to be pointed out that these modifications sometimes lead to distance “percentages” that are greater than 100.

## References

- Atkinson, Quentin., Andrew Meade, Chris Venditti, Simon Greenhill, and Mark Pagel (2008), ‘Languages evolve in punctuational bursts’. Science, 319, 588.
- Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman (2009), ‘Adding typology to lexicostatistics: a combined approach to language classification’. Linguistic Typology, 13, 167-179.
- Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai (2008), ‘Automated classification of the world’s languages: A description of the method and preliminary results’. STUF – Language Typology and Universals, 61, 285-308.
- Campbell, Lyle (2004), Historical Linguistics. An Introduction, 2<sup>nd</sup> ed. Edinburgh: Edinburgh University Press.
- Campbell, Lyle and William J Poser (2008), Language Classification. History and Method. Cambridge: Cambridge University Press.
- Dryer, Matthew S. (1989), ‘Large linguistic areas and language sampling’. Studies in Language, 13, 257-292.

- Dryer, Matthew S. (2005), 'Genealogical language list', in M. Haspelmath, M. Dryer, D. Gil, and B. Comrie (eds.), The World Atlas of Language Structures. Oxford: Oxford University Press, pp. 584-644.
- Dunn, Michael, Stephen C. Levinson, Eva Lindström, Geer Reesink and Angela Terrill (2008), 'Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia'. Language, 84, 710-759.
- Felsenstein, Joseph (2004), Inferring Phylogenies. Sunderland, Massachusetts: Sinauer Associates, Inc.
- Garrett, Andrew (2006), 'Convergence in the formation of Indo-European subgroups: Phylogeny and chronology', in P. Forster and C. Renfrew (eds.), Phylogenetic Methods and the Prehistory of Languages. Cambridge: McDonald Institute for Archaeological Research, pp. 139-151.
- Goh, Gwang-Yoon (2000), 'Probabilistic meaning of multiple matchings for language relationship'. Journal of Quantitative Linguistics, 7, 53-64.
- Greenberg, Joseph (1987), Language in the Americas. Stanford: Stanford University Press.
- Guy, J. B. M. (1980), Glottochronology without Cognate Recognition. Pacific Linguistics, Series B, No. 79. Canberra: Australian National University.
- Haas, Mary R. (1958), 'Algonkian-Ritwan: the end of a controversy'. International Journal of American Linguistics, 24, 159-173.
- Holman, Eric W. (2009), 'Do languages originate and become extinct at constant rates?' Paper presented at the Swadesh Centenary Conference, Max Planck Institute for Evolutionary Anthropology, Jan. 17-18, 2009

- Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker (2008a), 'Advances in automated language classification', in A. Arppe, K. Sinnemäki, and U. Nikanne (eds), Quantitative Investigations in Theoretical Linguistics. Helsinki: University of Helsinki, pp. 40-43.
- Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker (2008b), 'Explorations in automated lexicostatistics'. Folia Linguistica, 42, 331-354.
- Huelsenbeck, John P. and Fredrik Ronquist (2001), 'MRBAYES: Bayesian inference of phylogenetic trees'. Bioinformatics, 17, 754-755.
- Huson, Daniel H. and David Bryant (2006), 'Application of phylogenetic networks in evolutionary studies'. Molecular Biology and Evolution, 23, 254-267.
- Kessler, Brett (1995), 'Computational dialectology in Irish Gaelic', in Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers Inc., pp. 60-67.
- Kondrak, Grzegorz and Tarek Sherif (2006), 'Evaluation of several phonetic similarity algorithms on the task of cognate identification', in Proceedings of the COLING-ACL Workshop on Linguistic Distances. Sydney, Australia, July 2006, pp. 43-50.
- Levenshtein, Vladimir I. (1966), 'Binary codes capable of correcting deletions, insertions, and reversals'. Cybernetics and Control Theory, 10, 707-710.
- McMahon, April M. S. and Robert McMahon (2005), Language Classification by Numbers. New York: Oxford University Press.
- Müller, André, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Eric W. Holman, Dik Bakker, Oleg Belyaev, Dmitri Egorov, Robert

- Mailhammer, Anthony Grant, and Kofi Yakpo (2009), 'ASJP World Language Tree: Version 2 (April 2009)'. Available on the home page of the ASJP project: <<http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm>>.
- Nakhleh, Luay, Don Ringe, and Tandy Warnow (2005a), 'Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages'. Language, 81, 382-420.
- Nakhleh, Luay, Tandy Warnow, Don Ringe, and Steven N. Evans (2005b), 'A comparison of phylogenetic reconstruction methods on an Indo-European dataset'. Transactions of the Philological Society, 103, 171-192.
- Nerbonne, John, Wilbert Heeringa. and Peter Kleiweg (1999), 'Edit distance and dialect proximity', in D. Sankoff, David J. Kruskal (eds.), Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison. Stanford: CSLI, pp. v-xv.
- Nichols, Johanna (1996), 'The comparative method as heuristic', in M. Durie and M. Ross (eds.), The Comparative Method Reviewed. Regularity and Irregularity in Language Change. Oxford: Oxford University Press, pp. 39-71.
- Nichols, Johanna, and Tandy Warnow (2008), 'Tutorial on computational linguistic phylogeny'. Language and Linguistics Compass, 2, 760-820.
- Oswalt, Robert L. (1970), 'The detection of remote linguistic relationships'. Computer Studies in the Humanities and Verbal Behavior, 3, 117-129.
- Pinnow, Heinz-Jürgen (1959), Versuch einer historischen Lautlehre der Kharia-Sprache. Wiesbaden: Otto Harrassowitz.

- Saitou, Naruya and Masatoshi Nei (1987), 'The neighbor-joining method: a new method for reconstructing phylogenetic trees'. Molecular Biology and Evolution, 4, 406-425.
- Sapir, Edward (1913), 'Wiyot and Yurok, Algonkin languages of California'. American Anthropologist, 15, 617-746.
- Schleicher, August (1853), 'Die erstern Spaltungen des indogermanischen Urvolkes'. Allgemeine Monatsschrift fuer Sprachwissenschaft und Literatur, Sept. 1853: 786-787.
- Schmidt, Wilhelm (1906), 'Die Mon-Khmer-Völker, ein Bindeglied zwischen Völkern Zentralasiens und Austronesians'. Archiv für Anthropologie, 33, 59-109.
- Serva, Maurizio and Filippo Petroni (2008), 'Indo-European languages tree by Levenshtein distance'. Euophysics Letters, 81, paper 68005 (March 2008).  
[Online journal: <http://www.iop.org/EJ/journal/EPL>].
- Sokal, Robert R. and Charles D. Michener (1958), 'A statistical method for evaluating systematic relationships'. University of Kansas Science Bulletin, 38, 1409-1438.
- Swadesh, Morris (1950), 'Salish internal relationships'. International Journal of American Linguistics, 16, 157-167.
- Swadesh, Morris (1954), 'Perspectives and problems of Amerindian comparative linguistics'. Word, 10, 306-332.
- Wichmann, Søren, Eric W. Holman, Dik Bakker, and Cecil H. Brown (2009a), 'ASJP lexical similarity as a measure of language genetic relationship'.  
<<http://email.eva.mpg.de/~wichmann/ASJPPerformance14.doc>>.

Wichmann, Søren, Paulo Murilo Castro de Oliveira and The ASJP Consortium (2009b), 'Elbow effects: a universal feature of language taxonomies'. Presented at the Start-up-Meeting of the study group "Classification and Evolution in Biology, Linguistics and History of Science", Heinrich Heine University, Düsseldorf, June 11-12, 2009.

Wichmann, Søren and Arpiar Saunders (2007), 'How to use typological databases in historical linguistic research'. Diachronica, 24, 373-404.

Zide, Norman H. (ed.) (1966), Studies in Comparative Austroasiatic Linguistics. Indo-Iranian Monographs, 5. The Hague: Mouton.