Reconstructing a family tree (in abstracto)

Tamás Biró

Part 1: The simple story

Imagine, we have four languages: A, B, C and D. We have good reasons to suppose that they originate from a common ancestor, proto-ABCD, and we would like to draw a "family tree". The following table shows their properties for five different features, such as "what is the word for 'king'?" or "what is the 2nd person singular masculine suffix?" or "is there a definite pronoun?":

(1)

	A	B	C	D
F1	x	x	x	x
F2	y	z	z	z
F3	i	j	j	k
F4	f	g	g	h
F5	l	m	n	o

Looking at feature F1, we observe that all four languages share property x. So we can safely suppose that proto-ABCD also had that same property, which was subsequently inherited by all its descendents. As an example, proto-Semitic must have had two genders, for all Semitic languages have two (maybe with the exceptions of some modern dialects – to be checked). Yet, this observation does not help us categorizing our four languages into subgroups.

So let us turn to feature F2. Here we can see that languages B, C and D form a subgroup, as opposed to A. Features F3, F4 and F5 also confirm that language A stands alone, but they also help us in identifying smaller groups within the subgroup formed by B, C and D. Namely, B and C are more similar to each other than either of them is to D.

We can also quantify similarities of languages in terms of features they share. The similarity of language A to B is 1: they only agree in one feature. Likewise, the similarity of language A to either C or D is also 1. The similarity of D to either B or c is 2, while B and C share not less than four of the five features.

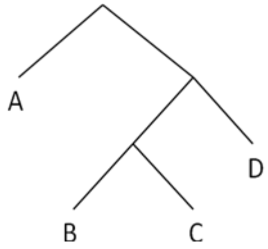
Let's draw the isoglosses in table (1) by using thick lines:

(2)

	A	B	C	D
F1	x	x	x	x
F2	y	z	z	z
F3	i	j	j	k
F4	f	g	g	h
F5	l	m	n	o

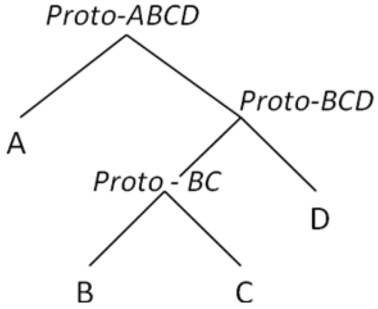
Imagine we would cut this table with a pair of scissors along these thick lines. With some imagination (and after moving the uppermost row with the names of the languages to the bottom of the table), we would obtain the following tree:

(3)



Please note that what we have done so far is mere clustering, grouping languages according to their similarities, without any reference to alleged historical developments. But this tree can also be looked at as a family tree:

(4)



Can it? In fact, in order to be convinced about such a family tree, we would like to reconstruct the common ancestors of languages A, B, C and D – at least to some degree. Furthermore, we also expect to have some “family history”. Otherwise, the nodes “Proto-ABCD”, “Proto-BCD” and “Proto-BC” remain nothing more than ad hoc names for the groups defined above.

Since B and C share the values x, z, j and g for the features F1, F2, F3 and F4, respectively, it makes sense to hypothesize that Proto-BC also shared these features. (Sometimes that’s not the case in historical linguistics, but then an argument must be made.) We cannot know, however, the value of F5 for this proto-language. It might have been m or n. Or even o, or maybe something else. In such a case, the historical linguists might remain agnostic, unless they have some additional arguments. For instance,

(a) A change m > n is probable (such as in the case of a simplification), while n > m is greatly improbable. Then the historical linguist will argue that Proto-BC must have had m, which was retained in language B, and which turned into n in language C.

(b) Neither m > n, nor n > m is a likely change. But one can suppose that a hypothetical *p in Proto-BC could easily be turned into m in language B, and into n in language C.

(c) Proto-BC has many descendents, B and C only being two of them. If most of these descendents share m, with the exception of language C, then it is likely that Proto-BC also had m, retained in most descendents, but not in C.

Similarly, one can also reconstruct feature F2 of Proto-BCD as z, but might have a harder time to argue for some value of features F3, F4 and F5. To summarize, we arrive at the following table:

(5)

	A	B	C	D	Proto-BC	Proto-BCD	Proto-ABCD
F1	x	x	x	x	*x	*x	*x
F2	y	z	z	z	*z	*z	*y *or* *z ?
F3	i	j	j	k	*j	*j *or* *k ?	*I *or* *j *or* *k ?
F4	f	g	g	h	*g	*g *or* *h ?	*f *or* *g *or* *h ?
F5	l	m	n	o	*m *or* *n ?	*m *or* *n *or* *o ?	*l *or* *m *or* *n *or* *o ?

The * symbols should constantly remind us that these values are not established by hard evidence; they are mere reconstructions.

The reconstructed story goes like this (as it would have been recounted by nineteenth century linguists in the time of romanticism): Once upon a time, there was a language. Let us call it Proto-ABCD. We hardly know anything about it, but most probably it had value x for feature F1. The speakers of Proto-ABCD were nomadic people… One morning they woke up and decided to split into two groups. One of them became the ancestors of those speaking A today, and the other group became the ancestors of those speaking B, C and D today. As the two groups separated, their speech also diverged throughout the centuries. They developed different ways of expressing features F2 to F5, but maintained value x for feature F1. Maybe there was also a third group, but no evidence of their speech has survived. A few centuries or millennia later, the population speaking Proto-BCD also split into two. Generations after the separation, one of the two subgroups developed language D (or an earlier stage thereof), while the other group developed Proto-BC. And so on.

These kinds of stories fit very well into nineteenth century romantic national (and nationalistic) historiographies.[1] But life is never so simple…

Part 2: When the story becomes more complicated

Now imagine, and this is what happens most often in historical linguistics, that table (1) is replaced by the following table (6). It is almost the same, but there is a slight change:

(6)

	A	B	C	D
F1	x	x	x	x
F2	y	z	z	z
F3	i	j	j	k
F4	h	g	g	h
F5	l	m	n	o

Unlike earlier, now A and D share the same value for F4. What are the consequences of this tiny fact? The consequence is that generations of linguists will debate how to categorize languages A, B, C and D.

[1] In some cases they did not. For instance, it was in the middle of the nineteenth century that linguists realized that the Hungarian language was related to Finnish, Lappish, Estonian, and to some other languages, mainly in distant Siberia. These newly discovered relatives were very poor, often hunter-gatherer or nomadic people, without any form of literacy. Even Finland was in those days an underdeveloped periphery of tsarist Russia, maybe one of the most despised countries in Europe. Thus, the discovery of historical linguistics did not fit into the romantic and glorious narrative of the Hungarian nation, and hurt the self-image of many Hungarian, provoking a rejection of the Finno-Ugric identification. In fact, this rejection is still very prominent in current chauvinistic discourse in Hungary.

There will be one fraction that will rely on F2 to draw the main borderline (isogloss), distinguishing between A, on one hand, and B, C and D, on the other. Yet, another group of linguists will argue that F4 must be used, and so languages A and D form one group, as opposed to group B and C. Features F1, F3 and F5 can be explained in both theories. The F2-ists will have a hard time explaining the similarity of A and D for feature F4, while the F4-ists will have to answer why D is similar to B and C for feature F2. Can we help them to come up with an account?

Suppose that there are further evidence (or further arguments) favoring the F2-ist position. Still, a number of explanations can be given for the similarity of A and D for feature F4.

First, it may be a case of coincidence. Whatever were the proto-languages, both A and D developed, independently of each other, the same value for F4. Second, some factors might have increased the likelihood of such a coincidence. If language typology (a survey of all human languages in the world) only allow for a few values for F4,[2] then any two randomly chosen languages will share the same value with a considerable chance. Cross-linguistic tendencies (universals) may also prefer certain values, further increasing the chances of coincidence.

Third, if the two languages otherwise share many other features, their structures are similar, while language use (culture, technological development, etc.) impose the similar constraints and demands, then the two languages are likely to develop similar "solutions" to these challenges. These developments are independent of each other, but triggered by similar factors. Irregularities inherited from the common ancestor or seemingly illogical details in the inherited paradigms also call for some remedies, which may be similar in otherwise geographically distant languages.

Fourth, it may very well be the case that the languages A and D influenced each other in a latter phase of their history. For instance, French heavily influenced English millennia after the ancestor language of Latin split from proto-Germanic. Similarly, Late Biblical Hebrew and Rabbinic (Mishnaic) Hebrew adopted several Akkadian words, and Modern Israeli Hebrew adopts Arabic loanwords.[3]

Finally, and most interestingly, a situation such as the one in table (6) may also occur if h was the value of feature F4 in the common ancestor, Proto-ABCD. This value was maintained even after the original proto-language split up, and still persists in the otherwise very remote languages, A and D. However, Proto-BC introduced an innovation, replacing value h with value g, and that is what we find in contemporary descendants of Proto-BC, namely in languages B and C. Such a state-of-affair is a rare case when we are capable of arguing for the feature value used by the Proto-Language: that is the form attested on both peripheries, but not attested anymore in the descendents of the language where the innovation took place. We do not have to remain agnostic about the proto-language, as it happened in (5).

Which of the five possibilities should I argue for? It depends on the actual values of the features, and on our own (often subjective) judgments based on our experience as a linguist.

[2] For instance, there are only two options for the word order within a noun phrase: the adjective either precedes or follows the noun. (A third possibility could be that their order is not fixed.) The number of genders in a language is either zero/one (no gender difference), or two (masculine vs. feminine; masculine-feminine vs. neuter), or three, very rarely four, unless the language employs a dozen *noun classes*.

[3] It is known that early historical linguistics (early nineteenth century), the idea of diverging features as a consequence of populations splitting up, inspired Charles Darwin to develop his theory of the evolution of the species. Later on, biological evolution theories have continuously influenced linguistics. Yet, here is a point where the analogy between the two disciplines must be given up. Namely, when a species is split in biology, leading to the emergence of two new species, then they will not crossbreed anymore.