Drawing a family tree based on isoglosses: dilemmas

Tamás Biró

1 Cases and definiteness

Hebrew has no case system, but it has a definite article *ha-* that precedes the noun. Aramaic has no case system either, and it marks the definiteness of a noun differently, with the suffix *-ā*. Arabic has a tripartite case system, and the definite article *(a)l-* is prefixed to the nouns. Consequently, Hebrew shares one feature with Aramaic, and one with Arabic, while the latter two share none of these two features.[1]

Which of the two languages is a closer relative of Hebrew? Based on different arguments, most scholars would consider Aramaic to be significantly more similar to Hebrew than Arabic. And yet, the bits of information just introduced do not necessarily confirm this view.

The following chart summarizes the picture:

	definite article at the front	definiteness at the end
cases	Arabic	–
no cases	Hebrew	Aramaic

Table 1: *Some Semitic languages*

Interestingly, Germanic languages also display a similar 2×2 feature table:

	definite article at the front	definiteness at the end
cases	German, Yiddish, etc.	Icelandic, Faroese, etc.
no cases	English, Dutch, etc.	Swedish, Danish, etc.

Table 2: *Some Germanic languages*

How should we draw a family tree for these languages, based on the information presented above?

[1] Arabic and Hebrew also share the *-na* suffix in the 2nd person plural forms of the prefix (imperfect) conjugation, as opposed to the *-ā(n)* found in Aramaic and further Semitic languages. This is the shared innovation used by Robert Hetzron (1976) to argue for grouping Hebrew and Arabic together, as we shall see it below.

2 From isoglosses to trees

We first have to draw **isoglosses**: lines (physical on a dialect map, or imaginary among the languages in our charts) that separate language varieties characterized with a phenomenon from those characterized by an alternative phenomenon. For instance, languages with a case system from languages without a case system. Or languages with definite articles from languages with definite suffixes. We shall say that an isogloss corresponds to a **feature**, which takes different **values** on the two sides of the isogloss.

Feature 'marking definiteness' can take, at least, the following three *values*: 'no marking', 'marking at the front', 'marking at the end'. The languages we are now considering make only use of the latter two, although closely related (older) languages would necessitate reference to the first value, as well. Similarly, the *feature* 'presence of a case system' can have two *values*, 'yes' and 'no'. Note that this is an instance of a *binary feature*, a feature with two possible values by definition. We could also have approached the phenomenon differently, distinguishing between many possibilities: having either zero, or two, or three, or four, etc. cases. For the sake of simplicity, we currently do not do so.

Let us consider the 'definiteness' feature first. The corresponding isoglosses run vertically in Tables 1 and 2. The languages in the left columns of the tables would define a group (appearing in the left half of the tree), and the right columns another one (under the right branch):

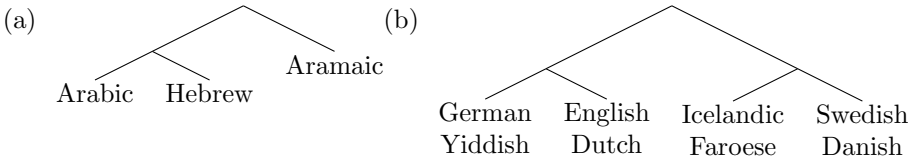

Figure 1: *Feature 'marking definiteness' as primary criterion for grouping languages.*

Within each main group of the Germanic languages, the presence or absence of a case system can still serve as a secondary criterion, defining sub-groups.

In turn, we could also use this second feature as the main criterion, defining the two major groups based on whether they have or they do not have cases. Now, the most important isogloss runs horizontally in Tables 1 and 2:

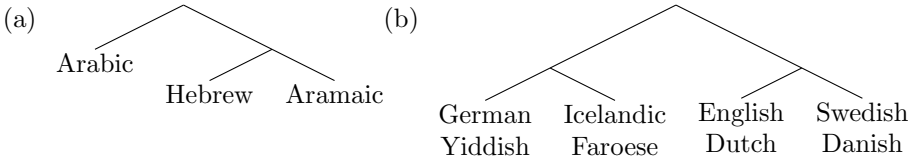

Figure 2: *Feature 'presence of cases' as primary criterion for grouping languages.*

3 Interpreting trees

Does Figure 1 or Figure 2 correspond better to our intuitive classification of these languages? Which of the trees is supported by further arguments? Most semitists would concur that Hebrew is closer to Aramaic than to Arabic, and so they would prefer Fig. 2. However, Fig. 1 is the one better describing the Germanic languages: the main divide separates the West Germanic group from the North Germanic (Scandinavian) one, and within each we find a Northern/Western and a Southern/Eastern subgroup.

However, we have to ask ourselves: what is the whole point of drawing trees? Some scholars are really compulsive with drawing trees. In fact, taxonomy trees have long been a standard way to organize our knowledge about basically any domain. Remember taxonomy in biology; remember library classification systems; remember the organization of any scientific book (such as this one) into chapters, sections and subsections. But does the visualization in the form of a tree really help us better understand the facts than, say, a table? Is a tree a better reflection of our knowledge than anything else?

Figure 1b hides an important observation: the further classification of the Western Germanic languages into subgroups have been done using the same criterion as the further classification of the Scandinavian languages. It is clear from the table, but not from the tree, that German and Yiddish are separated from Dutch and English by the same isogloss as Icelandic and Faroese are from Danish and Swedish. Thus, I would suggest, a table is simply a more efficient way of conveying complex information about both case systems and definiteness than a tree. In order to go for a tree nevertheless, further motivation is required.

One such motivation may be *having much more information*. For instance, German and Yiddish can be separated from Dutch and English based on several further criteria, such as the *High German consonant shift*,[2] the full presence of all three genders (Dutch only has two), the richness in diphthongs (German and Yiddish having some, English and Dutch having many more), and so on. The isoglosses drawn for many features (many different aspects of the languages) will fall together, English and Dutch lying on one side, and German and Yiddish on the other. Then, the classification of the Germanic languages by tree 1b efficiently condenses all this sea of information. It is plainly accidental that one of the many features separating the sub-branches of the Western Germanic languages coincides with one of the many features separating the sub-branches of the Scandinavian languages.

Another motivation requires a different perspective on a "family tree". A genetic tree is more than a (useful or less useful) **condensation of synchronic-typological information**. It is not only about the *cluster analysis*[3] of certain

[2] As an example of the /k/ > /x/ shift, contrast German *machen* and Yiddish *makhn* to English *make* and Dutch *maken*. For the /p/ > /f/ or /pf/ shift, compare German *Apfel* to Dutch *appel* and English *apple*; or German *Pferd* and Yiddish *ferd* for 'horse' to Dutch *paard*. Etymologically, English *up* and Dutch *op* are related to German *auf* and Yiddish *oyf*. Finally, the sound change /t/ > /s/ or /ts/ is exemplified by English *eat*, Dutch *eten* contrasted to German *essen* and Yiddish *esn*.

[3] *Cluster analysis* refers to a set of techniques developed in statistics seeking to group

feature values. Telling the truth, languages close to each other in a family tree may even be very different typologically. As a matter of fact, a family tree rather tells us a **narrative about the history** of these languages.

Put in over-simplistic terms, the tree of our three Semitic languages is about an ancient, hypothetical "proto-Arabo-Aramo-Hebraic" group splitting up. Did the ancestors of the Arabs leave the ancestors of the Arameans and Hebrews, or did the ancestors of the Arameans leave the ancestors of the Hebrews and Arabs? Based on the information we have, both 'scenarios' are equally probable... unless we also look at the *linguistic* content of the features and their values. Historical linguistics will help us break the impasse reached by the statistical approach.

The parent nodes in a historical tree are not just clusters of typologically or otherwise similar languages, but supposedly correspond to shared ancestral languages.[4] In the case of our three Semitic languages, and introducing proto-language names made up by myself:

Figure 3: *Possible proto-languages of Arabic, Hebrew and Aramaic.*

"Proto-HArabHaram" would be called by many Semitic studies scholars *Proto-Central Semitic*. They would refer to "Proto-HAram" as *Proto-Northwest Semitic*. However, what matters is not the names but the features of those postulated proto-languages.

Thus, Proto-HAram in 3a can be argued not to have a case system.[5] Similarly, Proto-HArab in 3b can be argued to have had a definite article (possibly **hal-*, with the /l/ tending to assimilate into the subsequent consonant). We remain agnostic regarding the article in Proto-HAram, and the case system in Proto-HArab. Proto-HArabHaram might or might not have a case system. If it did – as it most probably did, given the evidence from further Semitic languages – then it remained so in Arabic. Similarly, it might have had any value for marking definiteness. (Further arguments suggest it did not mark definiteness at all.) Finally, independent evidence prefers the narrative behind tree 3a.

However, the next sections will show that pure logic combined with general historical linguistics might lead to a very different conclusion.

---

observed data into clusters. Refer to `http://en.wikipedia.org/wiki/Cluster_analysis`. In particular, the task at hand is *hierarchical clustering*, which aims at building a hierarchy of clusters based on the data (`http://en.wikipedia.org/wiki/Hierarchical_clustering`).

[4] It is important to note that two languages having a common ancestor does not presuppose that the populations speaking them are also related ethnically and biologically. Populations undergoing *language shift* – such as the Bulgars, originally Turkic – illustrate the point.

[5] This sentence may be true in the context of the current train of thought. It will be reconsidered, however, when you learn more about the Northwest Semitic languages.

4 Searching for the most likely narrative

There are a few sciences, such as mathematics, in which claims are proved (or disproved) in an exact way. Most disciplines, however, are not so lucky. If you work in such a discipline, then the best you can do is to seek the *most probable* explanation of your data. Such has been the case for historical linguistics, too: implicitly until recently, and extremely explicitly – quantifying likelihood in terms of probability theory – by the so-called *phylogenetic methods*.

At this point, we may not ignore the linguistic content of the features and their values anymore. For instance, linguists have long observed that case systems disappear much more easily than they emerge from nothing. Case endings were gradually weakened and subsequently totally left out in many Semitic and Indo-European languages. By fixing the word order (such as having the subject precede, and the object follow the verb), and by introducing prepositions (to express a genitive construction or a dative object among others), these idioms developed alternative ways of expressing grammatical relations, and so case endings turned superfluous. This change may take place in a couple of hundreds of years, as opposed to the many thousand years it takes to develop a typical case system: Nouns are grammaticalized to form prepositions or postpositions;[6] which then gradually become shorter, cliticize, and subsequently turn into affixes;[7] and these affixes need to be reorganized into a system of cases.

It follows that it is much less likely for Arabic to have developed a case system from a proto-language not having it, than for Hebrew and Aramaic (together, or independently from each other) to have lost a case system. The most probable narrative begins with a "Proto-HArabAram" *with* cases. Without ample experience in historical linguistics, however, a pure statistical method might prefer a "Proto-HArabAram" *without* cases, since under this hypothesis only one of the three languages (*viz.* Arabic) needs to be postulated to have undergone a change (see, for instance, Fig. 4a below).

Compare the scenarios displayed on Fig. 4. Many more are possible: could you list all eight of them? Both proto-languages in each of the two trees on Fig. 3 can either have or not have a case system. Check that all four scenarios not depicted on Fig. 4 involve the *emergence* of a case system along the path from Proto-HArabAram to Arabic, corresponding to not a very probable narrative.

Let us return to the parenthetical remark "together, or independently from each other" two paragraphs earlier, and introduce another notion. The philosophical principle referred to as *Ockham's razor* (named after scholastic philosopher William of Occam, c. 1287–1347) prefers the hypothesis with the fewest assumptions, whenever alternative hypotheses are offered. A widely used idea in the philosophy of science, but also in machine learning, it has become the foundational postulate of phylogenetic methods: the most likely family tree is the *most parsimonious* one, the one with the least change along its edges.

[6] Think of *panim* in Hebrew *lifnei*. Think of *front* in English *in front of*. Think of *mell* in Hungarian *mellett*.

[7] "Feheruuaru rea meneh hodu utu rea", where Hungarian suffix *-ra/-re* still appears as a postposition.

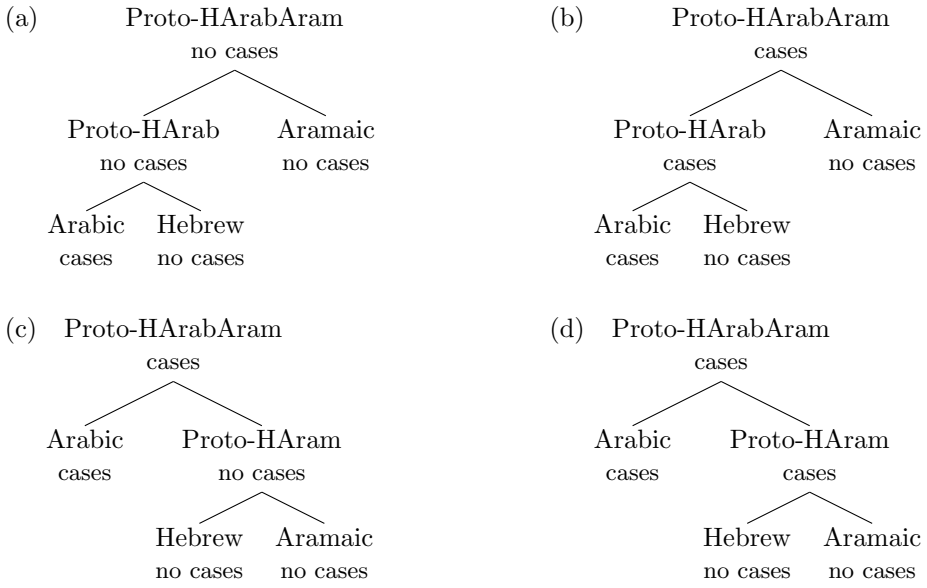

Figure 4: *Some scenarios for the development of case systems in the common ancestors of Arabic, Hebrew and Aramaic.*

Let us illustrate this principle with the disappearing case systems in Hebrew and Aramaic. Ockham's razor prefers assuming such a loss happening once to twice in the history of the languages under consideration. It is 'more parsimonious' to hypothesize that the Proto-HArabAram case system ceased to exist in Proto-HAram (tree 4c), and so it is missing from both daughters of this node, than to hypothesize independent developments in the two languages. The 'less parsimonious', hence less preferred alternative hypotheses include supposing that Proto-HAram *did* have cases (tree 4d); as well as positing the loss of cases on both right-branching edges of the tree 4b.

Such should be our line of argumentation, as long as we have not yet encountered Northwest Semitic varieties *with* a case system, such as Ugaritic and the Old Canaanite of the Amarna letters. But then, we are forced to reconsider our conclusion: "Proto-HAram", which was also the common ancestor of Ugaritic and Old Canaanite, did have cases. Luckily for us, losing a case system is really easy. Therefore, tree 4d with Proto-HAram having cases is an hypothesis that is hardly less parsimonious than the alternative tree 4c preferred thus far. In fact, including Ugaritic and Old Canaanite into the picture changes the game: Hebrew and Aramaic losing their case systems independently from each other is by far a more likely narrative than Ugaritic and Old Canaanite developing theirs from a case-less Proto-Northwest Semitic.

5 Hetzron's two principles of reconstruction (1): *Principle of Archaic Heterogeneity*

Decades before the advent of phylogenetic methods, Robert Hetzron (1976) offered a principled way to tackle a question such as ours. He offered two principles, one for determining the direction of a change, and one for weighting changes when defining language groups.

His *Principle of Archaic Heterogeneity* posits that it is more likely to move from a heterogeneous state to a homogeneous one, than vice versa:

> [...] when cognate systems (i.e. paradigms) in related languages are compared, the system that exhibits the most inner heterogeneity is likely to be the closest to the ancestor-system. (p. 89)

His example concerns the consonants of the suffix conjugation:

	1c Sg.	2m Sg.	2f Sg.
Akkadian	[k]	[t]	[t]
Geez	[k]	[k]	[k]
Arabic	[t]	[t]	[t]
Hebrew	[t]	[t]	[t]
Aramaic	[t]	[t]	[t]

Table 3: *Consonants of the suffixes in the suffix conjugation.*

Hetzron suggests to reconstruct the Proto-Semitic paradigm as *[k]/*[t]/*[t], as preserved in Akkadian. This reconstruction could have been supported by a comparison to Afro-Asiatic languages (such as the Egyptian *Old Perfective*), while the suffixes might be argued to originate in cliticized pronouns (**anāku*, **anta* and **anti*). Still, Hetzron brings a different argument: *Paradigmatic levelling* – elements of a paradigm turning more similar[8] – is a recurrent phenomenon in historical linguistics. Therefore, Hetzron argues, it is more probable for a heterogeneous paradigm to undergo paradigmatic levelling, than for elements in a homogeneous paradigm to become divergent.

A tripartite case system is more heterogeneous than having no cases. Hence, Hetzron's Principle of Archaic Heterogeneity can also be applied to our data: Proto-HArabAram most likely had three cases, as in our favorite trees 4c and 4d. Observe that we have come to this conclusion purely based on the data from Arabic, Hebrew and Aramaic, and without reference to corroborating evidence from Ugaritic, Old Canaanite and Akkadian.

[8] An example for *paradigmatic levelling* in the history of Hebrew suffix conjugation is Biblical Hebrew *kətabtém* replaced by *katábtem* in Modern Hebrew, which happened by analogy to the rest of the paradigm *katábti*, *katábta*, etc. (The accent denotes stress. Modern Hebrew does not have vowel length.)

6 Hetzron's two principles of reconstruction (2): *Principle of Shared Morpholexical Innovations*:

What about marking definiteness, also presented as part of our original data set at the beginning of this chapter? We have seen that they result in a different tree of our three Semitic languages. Moreover, definiteness data are more reliable in generating the tree of the Germanic languages, a surprising result in light of the analogous tree of the Semitic languages.

Should we conclude from the fact that tree 1b better describes the Germanic languages that definiteness is a more reliable isogloss? Probably not. However, Hetzron's second principle may be brought as an *a priori* argument for basing a classification on this feature, rather than on the case system feature. Hetzron explains his *Principle of Shared Morpholexical Innovations* thus:

> The most arbitrary elements of language are the phonetic shape of morphological and lexical items (the requirement of arbitrariness safeguards against possible developments due to general tendencies), and the phonetic shape of morphological items is the least likely to be borrowed (as against lexical items). (p. 89)

Therefore, Hetzron argues, the strongest possible argument for the genetic interconnectedness of two languages is them containing identical morpholexical items, typically affixes with the same form and meaning. To be more precise, the forms ought to be sufficiently similar, displaying the expected sound correspondences (unless we consider them skewed reflexes).[9] Similarly, slight differences in meanings (functions) should be tolerated, especially if these differences can be explained by other factors, such as the structure of the language.

Hetzron employs the form of the feminine plural ending in the prefix conjugation. Hebrew and Arabic have *-na/nā*, whereas Aramaic has *-ān*. The Aramaic suffix includes an originally separate *-n*, and therefore the Aramaic form can be seen as identical to Akkadian and Geez *-ā*. Hence, the suffix reconstructed for Proto-Semitic is also *-ā*. Two possible narratives should be considered following these observations, those on Figure 5. Can you come up with further scenarios?

The probability of Hebrew and Arabic developing a suffix with the same form and same function independently of each other is very low. Therefore, the *Principle of Shared Morphological Innovations* suggests that the development took place only once in the history of the Semitic languages, in the common ancestor of Hebrew and Arabic. However, Aramaic retained the old suffix, which observation led Hetzron to conclude that Aramaic seceded first from Hebrew and Arabic, as shown on Tree 5b.

A similar train of thought can also be applied to the marking of definiteness. Remember that definiteness proved to be a better isogloss than the presence of

[9] If the forms are too similar, not even differing in expected regular sound correspondences, then one should suspect borrowing.

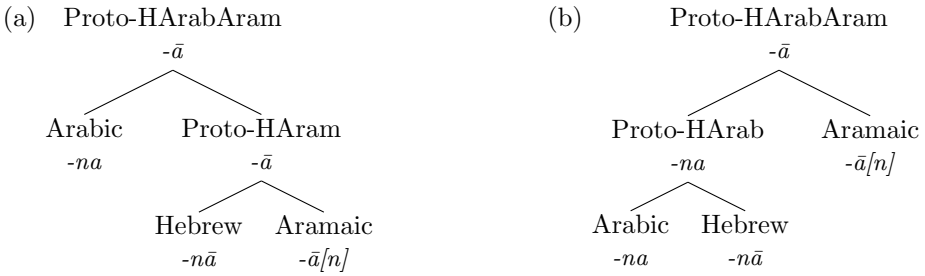

Figure 5: *Possible proto-languages of Arabic, Hebrew and Aramaic.*

a case system for the Germanic languages. This is not a coincidence. Proto-Germanic (as exemplified by the Gothic language) did not mark definiteness. Introducing a definite article preceding the nouns, and a definite suffix were both morphological innovations. Ockham's razor, parsimony and Hetzron's second principle point to the same direction: each innovation took place only once, in proto-West Germanic, and in proto-Scandinavian, respectively. Hence, our argument for tree 1b. The counter-argument, which would support tree 2b is much weaker, because a case system disappearing is not a morphological innovation that would fall under Hetzron's second principle. It is a much more likely scenario to postulate the introduction of a definite article once, the introduction of a definite suffix once, and the loss of the case system several times, than to posit the introduction of a definite article twice, the introduction of a definite suffix twice on the loss of the case system a single time.

In the case of our three Semitic languages, here are two scenarios:

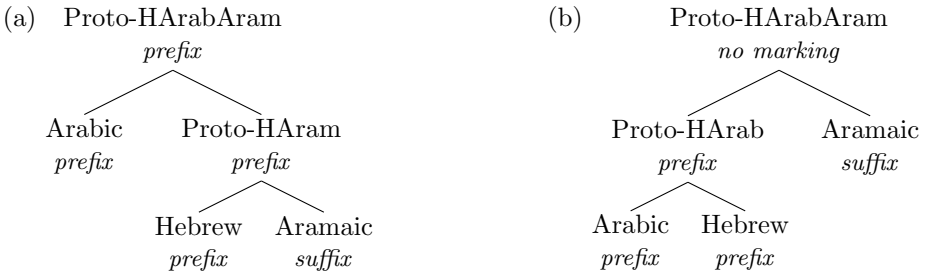

Figure 6: *Possible proto-languages of Arabic, Hebrew and Aramaic.*

Scenario 6a is based on a tree that is consistent with our prior preferences: Hebrew and Aramaic being grouped together. However, this tree only works if we posit a prefixed definiteness marker for Proto-HArabAram. Otherwise, we would have to hypothesize that Arabic and Hebrew *independently* developed a similar morpholexical item for the same function – a hypothesis that is dispreferred by Hetzron's second principle.

Scenario 6a, however, raises two problems. First, we have no independent argument to suppose a prefixed definiteness marker in Proto-HArabAram. The

lack of definiteness marker in Akkadian and Geez suggests no definiteness marker in Proto-Semitic; and even not in Proto-Northwest Semitic, for it is also absent from Ugaritic. Second, it is more probable to suppose that the Aramaic definite suffix (*status emphaticus*) developed in a language variety without any definiteness marker, than to suppose the transformation of a prefix into a suffix. Similarly, it is also preferable to suppose that Ugaritic reflects Proto-Northwest Semitic without a definiteness marker, than to suppose this marker to be lost between Proto-Northwest Semitic and Ugaritic.

Scenario 3b is therefore more plausible, parallelling the developments in the Germanic languages. The hypothetical proto-language had no way to mark the definiteness of a noun (or noun phrase), as mirrored by the earliest languages attested in the language family (Ugaritic here, Gothic there). Then, two morpholexical innovations took place: the common ancestor of one branch developed a determined article preceding the nouns, while the common ancestor of the other branch developed a suffix with the same function.

To summarize, employing Hetzron's *Principle of Shared Morpholexical Innovations* both to the 2nd person feminine plural suffix of the prefix conjugation, and to the marking of definiteness prefers grouping Hebrew and Arabic together, apart from Aramaic. What is wrong with this principle?

The answer is simple: it ignores areal effects, that is, the possibility of shared innovations in a language contact situation.

References

Hetzron, R. (1976). Two principles of genetic reconstruction. *Lingua*, *38*(2), 89–108.