philosophisch-historischen Classe der Kaiserlichen Akademie der Wissenschaften 134/7.

Meyer-Lübke, Wilhelm
1904 Albanisches. In: *Grundriß der romanischen Philologie*. I - II. Straßburg: Karl J. Trübner, 1888 - 1902.

Miklosich, Franz
1871 *Albanische Forschungen. II. Die romanischen Elemente im Albanischen.* Wien [= *Denkschrifte der Kaiserlichen Akademie der Wissenschaften* XX].

Orel, Vladimir
1996 *Albanian Etymological Dictionary*. Jerusalem[ms.].

Pokorny, Julius
1959 *Indogermanisches etymologisches Wörterbuch*. I. Bern: Francke.

Tagliavini, Carlo
1937 *L'albanese di Dalmazia.* Firenze [= *Biblioteca dell'Archivum Romanicum* XXII, ser. 2].

Trubachev, Oleg N. (ed.)
1975-1987 *Etimologicheskii slovar' slavianskikh iazykov* II - XIII. Moscow: Nauka.

Vasmer, Max
1921 *Studien zur albanesischen Wortforschung.* Dorpat: Universität Verlag.
1970 *Étimologicheskii slovar' russkogo iazyka* II. Moscow: Progress.

Macro Families: Can a Mistake Be Detected?

Ilia Peiros
The University of Melbourne

One of the linguistic courses I did at University about 25 years ago was a course of Nostratics taught by Aaron Dolgopolsky. At that time I found myself rejecting Nostratics completely, due to the paucity of my knowledge of the language families discussed: "Why should I accept this strange idea? Comparative Slavonic is much more convincing and I know the languages involved". I also could not understand why Dolgopolsky, an outstanding linguist and respected scholar spent his time on such an odd theory rather than doing something more convincing. A few years later, I joined the Nostratic Seminar, not from a love of Nostratics but because I liked the people involved: Dybo, Dolgopolsky, Starostin, Khelimsky and many others. Through this opportunity for observation and discussion, I reached some understanding of how the Nostraticists work, mainly what is the theoretical background of long-range comparisons. In this article I discuss only procedures for assessing claims that languages are related[1], and apply them to the study of East and Southeast Asian languages.

A hierarchy of least six different levels can be distinguished in comparative linguistics: Dialect, Language, Young Family, Developed Family, Old Family and Macro Family (cf. Jakhontov 1980). The distinction between these linguistic levels is not absolute, and in many cases it is hard to tell which level a particular genetic unit belongs to. Still, the distinctions are useful, and are worth including in further discussion. The units can not be defined in a formal way, but have some specific features:

• Dialect.

People who speak the same dialect understand each other without any restrictions. Differences in their speech varieties are mostly explained by the existence of sociolinguistic factors known to the speakers. Normally there are no doubts that a dialect represents a single genetic entity.

• Language.

It is important to distinguish the two notions, 'language' and 'sociolanguage'. Although two speakers of the same language may use different dialects with sometimes quite noticeable differences they will always understand each other if they discuss common topics. The criterion of mutual intelligibility is thus essential for the notion 'language'. However, two speech varieties are included in one sociolanguage if the speakers believe that they speak the same (socio)language, regardless of their actual ability to understand each other. Languages and sociolanguages form different combinations:

---

[1] The procedures are based on discussions which were conducted at the Nostratic Seminar during the 70s and 80s. Their formulation, however, took me another ten years, and I alone carry the responsibility for their current presentation.

One language – one sociolanguage

All speakers of Hungarian know that they use the same (socio)language. Slight differences in dialects do not prevent mutual intelligibility, which means that they belong to the same language.

One language – two or more sociolanguages.

This situation is represented, for example, by Serbian and Croatian: mutual intelligibility (one language), but people know that they speak different (socio)languages.

Several languages – one sociolanguage.

The Chinese 'dialects' present the best example of this type. It is well known that the differences between some of them are not less than between Slavonic languages and in many cases mutual intelligibility is not possible. On the other hand, speakers of Chinese 'dialects' know that they speak the same (socio)language and this fact often affects their behaviour.

In comparative linguistics the main focus is on languages, not sociolanguages.

People who speak different dialects of the same language can identify differences between their speech varieties and often like to discuss them. It is important that they talk about differences rather than similarities the existence of which do not surprise them at all. Genetic relationship of two dialects which belong to one language is always self evident for both speakers and linguists.

• Young Family.

Speakers of two languages which belong to a Young Family, for example Russian and Ukrainian, are able to communicate, but usually only with difficulty. The speakers can maintain conversation but it will often be interrupted by repetitions and changes. In many cases it is hard to say if we are dealing with two dialects of the same language or with two languages included in one Young Family. Genetic identity of the members of one Young Family is always clear for speakers and linguists. Speakers usually pay more attention to differences between their speech varieties rather than to similarities between them.

• Developed Families.

The Slavonic or Germanic languages provide examples of families which belong to this level. Normally, speakers of two languages which belong to the same Developed Family do not understand each other. However, they can find many similarities between any two members of such a family, say Russian and Czech, English and Danish, etc. Comparing such languages, a speaker would talk not about differences, which are taken for granted, but about similarities between them. Speakers' explanations of such similarities may vary, but quite often they are based on an assumption of common origin of the languages.

For comparative linguists, genetic relationship of members of Developed Families is usually self evident and does not cause any problems apart from the questions of classification. Data from Developed Families is quite transparent, and normally it is not too difficult to connect a reconstructed proto-form with its reflexes in recorded languages.

Most comparative linguists conduct their research at the level of Young or Developed Families. Quite often such linguists have native or near native command of several languages of the family, which allows them to operate with high quality first-hand data.

A reconstruction of a Developed Family's proto-language is based on comparison between spoken or written languages:

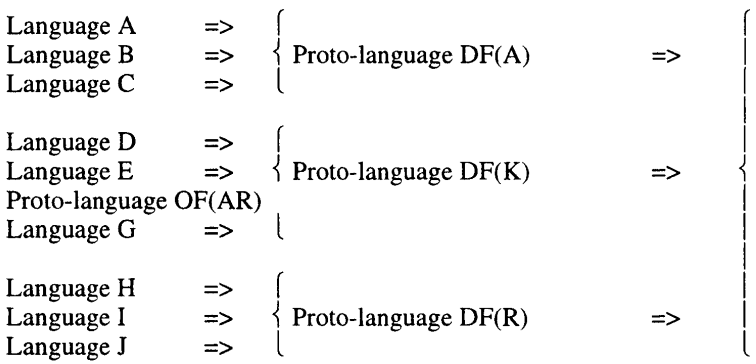
Language A =>
Language B => Proto-language DF
Language C =>

Sometimes, however, an intermediate reconstruction is needed if a Developed Family includes a Young Family as one of its branches.

• Old Families.

A good example of an old language family is Indo-European, which includes, among many other languages, French and Hindi. French speakers who study Hindi normally cannot identify forms of common IE origin, retained in both languages. Various changes which took place in the history of these languages have resulted in originally similar forms becoming absolutely different. For this reason, speakers cannot create any reasonable hypothesis about the genetic affiliation of modern languages which belong to different branches of an Old Family: their intuition does not work at this level. For comparativists, on the contrary, it is quite normal to deal with Old Families and their validity as families is widely accepted. It is important, however, that the existence of the IE family has been discovered not through comparison of modern languages like French and Hindi, but through knowledge of ancient languages like Latin, Greek and Sanskrit (which are much closer to each other and, I think, actually belong to the same Developed Family, as the similarities between them are easy to detect).

The Proto-IE language was originally reconstructed mainly through direct comparison of ancient languages, which meant that the work occurred at the level of Developed and not Old Families. Evidence from Proto-Germanic, Proto-Slavonic, Proto-Celtic, and other studies was subsequently incorporated into Proto-IE investigation, which nowadays is reoriented toward comparisons of reconstructed Proto-languages rather than simply archaic languages:



Theoretically, it is clear that intermediate proto-languages rather than recorded ones should be used in the reconstruction of Old Families. Each intermediate proto-language is associated with a Young or Developed family, which includes transparently related languages. If ancient and archaic languages are not known, success in the study of an Old Family depends on the existence of such intermediate reconstructions.

To do research at the level of an Old Family a linguist needs to be familiar, not only with the main languages of the family, but also with problems in the reconstructions of all proto-languages constituting the family. The amount of information required is many times more than in a study of a Developed Family, and this automatically reduces the number of specialists in Old Families. I do not know the exact number, but I estimate it to be roughly ten to twenty times less than the number of people studying Young or Developed families.

At the level of Old Families we face for the first time a conflict between the judgments of speakers and linguists: the latter recognise the genetic relationship, while the former do not.

• Macro families.

Comparing even very archaic languages which belong to different branches of a Macro Family, we will normally find only a limited number of similarities, which often are not fully convincing. That is why the justification of macro-families such as Illich-Svitych's Nostratics, and Starostin's Sino-Caucasian, is based not on comparisons of recorded languages, but on reconstructions of proto-languages constituting them, which in their turn can be based on reconstructions of younger proto-languages. A Macro Family reconstruction is therefore based on several levels of intermediate reconstruction conducted independently for each of its daughter proto-languages. Three or four levels of reconstruction are quite normal at this level of complexity:

I

Language A => ⎫
Language B => ⎬ Proto-language DF(AC)
Language C => ⎭

Language D => ⎫
Language E => ⎬ Proto-language DF(DG)
Language G => ⎭

Language H => ⎫
Language I => ⎬ Proto-language DF(HJ)
Language J => ⎭

Language K => ⎫
Language L => ⎬ Proto-language DF(KM)
Language M => ⎭

Language N => ⎫
Language O => ⎬ Proto-language DF(NP)
Language P => ⎭

Language Q => ⎫
Language R => ⎬ Proto-language DF(QS)
Language S => ⎭

Language T => ⎫
Language U => ⎬ Proto-language DF(TV)
Language V => ⎭

Language W => ⎫
Language X => ⎬ Proto-language DF(WY)
Language Y => ⎭

II

Proto-language DF(AC) => ⎫
Proto-language DF(DG) => ⎬ Proto-language OF(AH)
Proto-language DF(HJ) => ⎭

Proto-language DF(KM) => ⎫
Proto-language DF(NP) => ⎬ Proto-language OF(KQ)
Proto-language DF(QS) => ⎭

Proto-language DF(TV) => ⎫
Proto-language DF(WY) => ⎭ Proto-language OF(TW)

III

Proto-language OF(AH) => ⎫
Proto-language OF(KQ) => ⎬ Proto-language MF (AW)
Proto-language OF(TW) => ⎭

The Study of Macro Families is characterised by:

(i) the absence of archaic languages which could be directly compared to justify the relationship. Reliable similarities are found beetween reconstructed proto-languages, rather than between any recorded ones.

(ii) lack of transparency in reconstructions: normally one cannot tell whether recorded forms can be traced back to the proposed Macro Family reconstruction. To check this one needs to know the histories of all the daughter-families.

(iii) usually daughter-families are investigated by separate branches of comparative linguistics, with no significant tradition of cross reference.

These three features make any research at the Macro Family level extremely difficult. Because of the huge amount of data required, only people of exceptional ability can work at this level of complexity. I believe that altogether less than 50 scholars in the linguistic world can successfully study Macro Families. This gives them a brilliant opportunity to talk about their hypotheses and problems in an exclusive club ignoring the needs of other linguists. Should we simply wait for revelations issued from this club, or can we investigate their claims ourselves?

There are four major classes of objections to hypotheses about Macro Families:

(i) One can reject the whole idea of long range comparisons as the product of pure imagination and thus beyond true scholarship. Different arguments have been produced to support such an approach: "the comparative method cannot be applied to such remote periods of time", "the languages must have been different in the remote past and thus they would have followed other rules of

development", and many others of this type.[2] I think, however, that the underlying sense of all such claims can be formulated as follows: "I can not accept Nostratics or the Sino-Caucasian hypothesis as they contradict my intuition and they are too complex for me to evaluate. I do not have time to struggle with the evidence presented, and instead I will produce some general suggestions to support my feelings". Such claims are not, however, based on any linguistic evidence, and therefore we cannot seriously discussed them.

Leaving aside such general rejections two approaches are related to studies of particular Macro Families.

(ii) One can reject a Macrofamily hypothesis on the grounds that it does not fit with the broader prehistoric picture: "Nostratics is wrong, as it is not supported by extra-linguistic evidence", "I cannot fit the Sino-Caucasian theory into my understanding of Asian prehistory", and so on. It is clear, however, that such considerations can be used in the historical interpretation of linguistic hypotheses, but are not applicable to the discussion of these hypotheses within linguistics.

(iii) More serious objections come from the following direction. Talking about a Macro Family, a specialist in history of a particular Old Family conducts a thorough investigation of relevant reconstructions used in justification of the Macro Family. Taking, say Indo-European data in Nostratics, the linguist finds several wrong or unconvincing IE proto-forms included in Nostratic etymologies. This fact leads to the conclusion the whole Nostratic hypothesis "remains as yet a house of cards" (Vine 1991:31). Logically speaking, this is one of two possible ways to reject a Macro Family claim. To use this option properly, however, one needs to demonstrate that all or most Indo-European comparisons used in Nostratics are wrong.[3] If there are only a small number of incorrect Indo-European forms, they can be removed from Nostratic etymologies without destroying the whole hypothesis.

(iv) Another way to reject a Macro Family hypothesis is based on the analysis of methods used in its justification. It could be argued that the study of Macro Families belongs to comparative linguistics, and the same methods and procedures should be used in investigation of any language families: Young, Old or Macro. If we could demonstrate that the methods of comparative linguistics were violated, that would immediately take the corresponding hypothesis out of the discussion: all statements of comparativists have their meanings only within comparative linguistics.[4]

It seems to me, that the method of comparative linguistics provides us with reliable tools for the formal evaluation of any claims about genetic

[2] A example of such reasoning is represented by the following remark of an anonymous internal referee of Peiros, to appear: "Peiros believes that the step from Proto-Indo-European to "Proto-Nostratic" is just as straightforward as, eg. from Proto-Indo-Iranian to PIE. Does not seem to realise that after a certain period in time (ca. 10,000 years B.P. at a very generous estimate) genetic relationship is indistinguishable from borrowing or pure chance."

[3] It seems that this can be done within Japanese / Austro-Tai hypotheses for its Japanese and Miao-Yao components (see below).

[4] Note, however, that comparative linguistics often shares terminology with other theories, which use totally different methods.

relationship of languages and thus for evaluation of the validity of claims about Macro Families. The comparative method, as it can be described now, is based primary on the study of rather transparently related languages which usually belong to Young or Developed families. Old families are also studied at a level of transparency, but reconstructed instead of recorded forms are used. If we want to study even more ancient genetic units - Macro Families - we also need to operate with transparent relationship, which means that only a study based on reconstructed potential siblings is acceptable. In other words, to deal with a Macro Family we need to base our arguments on constituent proto-languages which reveal simple and transparent relationships. The collected data should be able to convince linguists who are not specialists in this particular theory, and who play here the same role as untrained speakers at the level of Young and Developed families.

Taking at random two languages A and B we could reach one of three possible conclusions as to their genetic relatedness:

(i) they are genetically related and share the same ancestor. Their relationship may be transparent as on the level of Young or Developed Families, or it could be more obscure, as between members of major branches of Old and Macro families;

(ii) language A is a direct or remote ancestor of B;

(iii) it is not known if A and B are genetically related. In such cases linguists would usually say that the languages are not genetically related, despite the fact that within comparative linguistics it is impossible to demonstrate the absence of genetic relationship. The only theoretically correct claim which can be made is that there is no available evidence that the two languages are related.

Let us limit our discussion to the first possibility: genetically related languages. It is generally accepted that languages are genetically related if they can be traced back to the same common ancestor. This means that strictly speaking if we want to demonstrate that languages A and B are related, we must present their ancestor, language C. With few exceptions C would be a proto-language, whose system is reconstructed through the comparative method based on a comparison of its daughter-languages. This leads us to a vicious circle: to prove that the languages are related we need a reconstructed proto-language, but to reconstruct it we need to know which languages are related. To overcome this contradiction we can use a working definition of genetic relatedness which does not include the notion of proto-language.

Related languages usually contain certain similarities which are traces of their common origin. Such similarities can be functional and / or material. In the case of pure functional similarities, certain parts of linguistic systems are organised similarly. For example, two languages might distinguish identical sets of noun classes, although the grammatical morphemes used to mark the classes could be quite different. Systemic features and their particular combinations do not appear at random, so it is not impossible that these similarities indicate genetic relationship. It is, also highly probable, however, that they are results of areal influences, typological universals and other non genetic factors. For this reason, functional similarities should never be used as the sole piece of evidence for genetic relationship and, in fact, they are not used as such in any well attested case.

The main body of evidence comes from material similarities. These include similarities between morphemes of the languages sometimes together

with similar irregularities found in the languages, such as between English and German irregular verbs: *drink, drank, drunk* vs. *trinken, trank, getrunken*. It is very hard to believe that such irregularities can be borrowed or result from independent development, so they are convincing indicators of possible genetic relationship. Unfortunately, however, they are not found often enough and genetic claims are primarily based on morphemic similarities and conclusions drawn from them. A list of similar morphemes found in the languages under investigation is absolutely crucial, and without it no genetic claims can be substantiated within comparative linguistics.

Before taking the next step in our discussion we need to clarify several notions. If the history and relationships between the languages are known:

• morphemes are called genetically related if they all result from the direct and uninterrupted development of the same morpheme of the proto-language. This morpheme is called their *proto-morpheme (proto-form).*

• morphemes which can be traced back to the same proto-morpheme are called *cognates.*

• a set of cognates developed from a single proto-form is called an *etymology.* An etymology thus includes only genetically related forms found in the languages under investigation.

• morpheme a in language A is a *reflex* of the proto-morpheme *f, if a is a result of direct development of *f in the history of A; phoneme a in language A is a *reflex* of the proto-phoneme *f, if a is a result of direct development of *f in the history of A.

If the history and interrelationship between the languages is not yet known:

• similar morphemes in those languages are called *resemblances.* There are various reasons why the morphemes may be similar: they could be cognates, borrowings, or even chance similarities.

•A set of resemblances found in the languages is called a *comparison.* No substantial claims can be made about the origins of a comparison. An etymology is a particular type of comparison, one which includes only genetically related morphemes.

Now using the notions of an etymology (= a set of genetically related cognates) and a comparison (= a set of resemblances which are not necessary genetically related) we can suggest a working definition of genetic relationship. Languages are genetically related if:[5]

(i) there is a sufficient number of comparisons consisting of resemblances found in these languages;

(ii) it can be demonstrated that these comparisons are etymologies in the strict sense and not borrowings or chance similarities. As the only accepted way to demonstrate the genetic nature of a comparison is to show that its resemblances are connected by systemic phonological correspondences (reflections of certain features of the proto-language phonological system), a list of systematic phonological correspondences is another necessary element for proof of genetic

[5] This definition does not specifically require identity of grammatical morphemes. It is based on my experience in comparative study of South-East Asian languages, which usually do not have developed grammatical systems, but still obviously form clear-cut genetic units.

relationship. In many cases the systematic correspondences also help us to identify loans.

Fulfilling these two conditions for the demonstration of genetic relationship provides us with information sufficient for phonological and lexical reconstruction of the proto-language. For the families with old morphology we should also be able to reconstruct common grammatical morphemes on the basis of comparisons between daughter language grammatical morphemes. According to the given definition, however, reconstructions are not required for proof of genetic relationship.

This definition is designed to work in the cases of transparent genetic relationships, like those represented in Young and Developed families, which are supported by the intuition of both speakers and linguists. It does not work directly for Old or Macro families, where instead of modern languages, one must work with their archaic ancestors, recorded or reconstructed. In research at this level it is still important to operate with rather transparently related languages and to apply the same two conditions, treating these (proto-)languages in the same way as modern ones.

In this paper I want to discuss the theoretical validity of evidence presented to support the following genetic claims: Sino-Caucasian, Japanese / Austro-Tai, Sino-Austronesian and Miao-Yao-Austroasiatic. The Sino-Caucasian theory (Starostin 1982, 1984) claims that three language families, Northern Caucasian, Eniseian and Sino-Tibetan are genetically related. The Japanese / Austro-Tai (JAT) theory claims the Austronesian languages are related to Kadai, Miao-Yao and Japanese (Benedict 1990). According to the Sino-Austronesian theory, Chinese and the Austronesian languages are genetically related (Sagart 1993; 1994), which contradicts both the SC and the JAT hypotheses. The Miao-Yao / Austroasiatic claim (Jakhontov 1981, Peiros to appear) connects Miao-Yao with the Austroasiatic rather than with Kadai and Austronesian families.[6]

Based on the requirements outlined in the working definition above, we can subdivide these genetic claims into three groups:

1. well supported claims: a sufficient number of comparisons, connected by systematic phonological correspondences are presented. Whether the genetic relationship is established depends on the quality of data presented, but the formal requirements of the definition are fulfilled. Starostin's SC theory belongs here: a list of comparisons is given, and major phonological correspondences are established. Strictly speaking, only claims of this type can be fully discussed and formally evaluated within comparative methodology.

2. plausible claims: these are based on a certain number of comparisons, but no systematic correspondences are established. Such claims are often just indications that further research is needed to 'upgrade' their level with more similarities and a set of phonological correspondences. I believe that the Kadai - Austronesian and Miao-Yao - Austroasiatic hypotheses belong here.

3. claims not supported by convincing evidence: the comparisons given do not necessarily indicate genetic relationship and no set of phonological

[6] I argue (to appear) that Kadai-Austronesian and Miao-Yao-Austroasiatic are two main branches of the Austric macrofamily.

correspondences is provided. I think that the treatment of Miao-Yao and Japanese in JAT and the whole Sino-Austronesian hypothesis belong here.

In evaluating genetic claims, we need to consider the following issues:

1. initial data: which languages in which form are compared;
2. how comparisons are identified;
3. quality of phonological correspondences, if any.

Initial data.

Two types of data can be used in justifying a genetic claim: reconstructed proto-forms and morphemes taken from recorded languages. At least five types of proto-forms can be found in the literature: real reconstructions, areal reconstructions, reflections, pre-reconstructions and ghost-reconstructions (This rather vague terminology is mine). The most reliable are *real reconstructions*. They are obtained through the strict universal procedure of comparative linguistics: (i) their identification is based on the system of phonological correspondences and plausible semantic relationships, (ii) their reflexes are found in all or major languages of the family and (iii) they can be definitely attributed to the proto-language level.

Sometimes a reconstruction is based on forms found in several languages and is confirmed by proper phonological correspondences, but it cannot be demonstrated that the form should be attributed to the proto-language level, rather than to a later period of the family's history. In such cases we are dealing with an *areal reconstruction* which could belong either to the proto-language of the whole family, to one of its daughter proto-languages, or represents unidentified areal influences on some languages of the family. The status of areal proto-forms is similar to that of proto-forms reconstructed for different branches of the family: in both cases we do not know at exactly what level of relationship they represent. Their usage undermines the validity of a genetic claim.

Many proto-forms used in the justification of JAT are areal reconstructions. Among them are all PAN forms based solely on Formosan data. The AN languages of Taiwan reveal similarities with Kadai and Japanese, which are not found in AN languages elsewhere. There are two possible explanation for this. One can assume that the Formosan languages have retained a significant number of PAN forms lost in other languages of the family (this is the position of Benedict), or it can be suggested that these languages preserve traces of contact with Kadai and/ or Japanese which took place after the disintegration of PAN (the geographical position of Taiwan makes this suggestion rather convincing). As we do not have enough data to choose between the two options, it would be better not to use these areal reconstructions in justifying this genetic claim.

If a morpheme is recorded in a language with a known history, but cognates are not found in other related languages, a linguist who believes that this morpheme is not a borrowing can assume that its ancestor form was also represented in the proto-language, and a corresponding proto-form can be reconstructed. Such "*reflections*" have less convincing power than real reconstructions, as there are no general reasons why they should be attributed to the proto-language level rather than to the level of one of its daughter (proto-) languages.

Reflections are often used in JAT: PAN **ʔumuq* 'pus' based only on Paiwan *umuq* (Benedict 1990: 232) is an example (if the phonological relationship between the Paiwan and Proto AN forms can be accepted). Obviously it is very hard to demonstrate that a reflection really belongs to the proto-language and is not, say, a later borrowing. Reflections, however, can be used in comparisons, providing they do not form the major body of evidence.

Two other types of proto-forms found in the literature, pre-reconstructions and ghost-reconstructions, do not, strictly speaking, belong to comparative linguistics. *Pre-reconstructions* are not based on a proper set of phonological correspondences but only on the intuition of the linguist who introduced them. Working with language families for which the comparative phonology is not well known a linguist may bring together fragments of historical information to gain an idea of how a proto-form might look. To transfer such pre-reconstructions into real reconstructions a detailed comparative phonology of the language family under consideration is needed. Without it, any genetic claim based on similarities between pre-reconstructions remains only a hypothesis.

The degree to which pre-reconstructions are convincing depends on such factors as:

(i) development of comparative phonology. For example, the main features of the PAN phonological system are known, but additional study is needed to work out detailed histories of its daughter families and their constituent languages. Any PAN form which is based on new data from languages with relatively obscure phonology, like Tsouic or Atayal, remains a pre-reconstruction (although perhaps a plausible one).

(ii) similarity between (proto-)languages. Forms of different Kadai branches are often rather similar, which in some cases makes pre-reconstructions quite convincing.[7] By contrast, relationships between Miao-Yao languages are much more obscure and the Proto-Miao-Yao pre-reconstructions used by Benedict are consequently less reliable.[8]

A *ghost-reconstruction* is the most treacherous type of proto-form found in the literature. It is usually based on a single morpheme, sometimes marginally represented in a language or simply on a mistake due to poor knowledge of the language's history. Four different Proto Miao-Yao ghost-reconstructions, each supported by a single form from only one Yao dialect can be found for example in the JAT comparison HOLD/ BITE/ CHEW (Benedict 1990: 209-211): **khamgam*^{B} (< Haininh Mun Yao *khamŋam*^{B} 'jaws'), **ŋgam*^{C} (< Haininh Mun Yao *gam*^{C} 'press with the hand, crutch'), **gom*^{C} (< Chianrai Mien Yao *kom*^{L} 'to fetter, shackle') and **ŋgom*^{A} (< Haininh Mun Yao *geom*^{A} 'hold in mouth'). No conclusions can be made on the basis of such ghosts.

Distinguishing these five types of proto-forms allows us to describe genetic claims as comparative (based primarily on true reconstructions) or

---

[7] In fact, the relations between Kadai languages are more complicated than they seem at first, and to work out a Proto-Kadai phonological reconstruction is a challenge (Peiros, to appear).

[8] My Miao-Yao reconstructions are based on a set of systematic correspondences between Proto Miao and Proto-Yao and forms represented in both branches of this family (Peiros, to appear).

heuristic (based on pre-reconstructions). Only comparative claims can formally justify a genetic relationship. Starostin's SC theory exemplifies the first type of claim, the other hypotheses mentioned above are all heuristic rather than comparative.

It is absolutely clear that the reconstructions used in a comparison should be self-reliant, which means that they should be obtained independently from each other, and not 'tuned' for better similarity. If a new version of a reconstruction is used only to support a genetic claim, we should be quite suspicious: often it means that the proto-forms have been 'tuned'. The most secure cases are when proto-forms are taken from already existing sources, comparative dictionaries or reconstructions made beforehand,[9] rather than suggested in the publication which makes the genetic claim. It seems to me that a priori a genetic claim based on previously known proto-forms is much more convincing than a claim based on proto-forms created especially for the purposes of its justification. That is why I still believe that the Benedict's original AT article (1942) is more convincing than his whole AT book (1975). Real reconstructions, by their nature, can not be 'tuned'; this is the 'privilege' of pre-reconstructions and ghosts.

In contrast to proto-forms, most recorded morphemes and words are real and reliable. In many languages, like Chinese or Japanese, due to various losses and mergers a modern form could be traced back to many different ancient forms. Only thorough investigation of the language's history can reveal its real ancestor. Such investigation is usually based on detailed study of historical phonology and lexicology. That is why I personally always have strong suspicions when external comparisons are based on a new version of the historical phonology of a language. Much more reliable are forms taken from historical dictionaries or phonological studies, rather than those adjusted for external comparisons. Only in the first case can one be sure that the forms are reconstructed properly. That is why it is appropriate to treat the Japanese forms in JAT with suspicion: the sources of Old Japanese forms in JAT often remain to obscure and in many cases are not supported by the history of Japanese (Vovin 1994: 373-376). In SC theory, however, all Archaic Chinese forms are taken from Starostin's Archaic Chinese reconstruction, completed much earlier than the SC studies began.[10].

In order to fully illustrate the effect that quality of initial data has on whether a genetic claim is convincing or not, it is worth comparing in detail Starostin's SC and Benedict's JAT theories.

The SC theory is based on the comparison of Proto North-Caucasian, Proto Eniseian and Proto Sino-Tibetan reconstructions which have been made absolutely independently from each other. In reconstructing Proto NC and Proto EN, the precise method of comparative linguistics was used including step by step movement from recorded languages towards their common ancestor. Several intermediate reconstructions, such as Proto-Lezghinian and Proto-

[9] The SC theory was originally discussed in the article which also included the Proto-Eneseian reconstruction. This reconstruction is, however, self-sufficient and is not based on data from other language families.

[10] The reconstruction has been published in 1989 (Starostin 1989), but it was completed much earlier.

Dagestanian, were created before dealing with Proto NC. Each of these reconstructions is based on a set of systematic phonological correspondences and a representative list of etymologies. An NC comparative dictionary is now published (Nikolaev and Starostin 1995). Proto-EN etymologies are given in the first part of Starostin 1982, with data demonstrating that the proto-forms are based on phonological correspondences and are not 'tuned'.

The situation with ST reconstructions is more complicated. When Starostin published his SC comparisons, the following ST data was available to him: Benedict's Tibeto-Burman reconstructions, Starostin's own Archaic Chinese reconstruction and our comparative ST (Peiros and Starostin 1996) dictionary which wa unpublished in that time. To make his SC results more convincing, Starostin chose to use Benedict's proto-forms together with his own Archaic Chinese forms, rather than quote from the unpublished dictionary. Data from the dictionary was used indirectly: only comparisons accepted in it were included in the SC etymologies. This approach, however, had a weakening effect on the whole theory, as Benedict's proto-forms are not true reconstructions supported by a complete set of systematic phonological correspondences but only pre-reconstructions reflecting Benedict's historical guesses.

The proto-forms used for justification of the JAT hypothesis are of quite different nature. To illustrate this, we can draw examples from Benedict's treatment of each of the four language families involved: Kadai, Miao-Yao, Austronesian, and Japanese.

The Kadai family includes 6 to 8 branches, with proto-languages reconstructable for at least three of them (Zhuang-Tai, Kam-Sui and Li). A comparison of these proto-languages with genetically isolated Ong Be and Likkja leads to Proto-Kadai, the reconstruction of which is not yet published (Peiros, to appear). 'Tuned' proto-forms from three intermediate reconstructions – Zhuang-Tai (or Tai) by Li Fangkuei (1977), Kam-Sui (Thurgood 1988) and Proto-Li (Matisoff 1988) – are used in Benedict's 1990 book. A good example is Benedict's Proto-Kadai form 'sugarcane' **[t]o[b]oi> *[t]o[w]oy > *$C_t$ooy^B* based on Zhuang-Tai **ʔooy^B* 'sugarcane', Kam-Sui **ʔooy^B* 'sugarcane' and Southern Li (which dialect?) *oi^C* 'maize' (1990, 232). It is based simply on obvious similarity between Zhuang-Tai and Kam-Sui forms and the need to connect them with Proto Austronesian **təbus* 'sugarcane'. This, and all other Kadai proto-forms discussed by Benedict, remain to be pre-reconstructions. Most Kadai languages are quite similar to each other (they perhaps form a Developed Family), and usually it is not too difficult to identify comparisons. However, intensive internal contacts and impact from Chinese, Vietnamese, Khmer and other Southeast Asian languages necessitate a detailed knowledge of Kadai comparative phonology for proper genetic interpretation of comparisons found.

The Miao-Yao family with its two main branches (Miao and Yao) presents another type of problem. Due to the occurrence of significant phonetic changes, identification of comparisons even in closely related Miao languages can be quite challenging, especially if forms are not known from more archaic dialects (languages). Phonological correspondences connect the main Miao dialects (Wang 1985), but a detailed Miao comparative dictionary does not exist. Proto-Yao is known mainly thanks to Purnell's reconstruction (1970), which requires some revision (Peiros, to appear), with extensive reliable data available for only one dialect (Lombard 1968). These limitations mean that Benedict's MY

proto-forms remain pre-reconstructions, less convincing than those suggested for Proto-Kadai. Their reliability is also undermined by possible Chinese borrowings which are hard to detect without proper phonological information.

The Austronesian family includes many hundreds of languages, grouped into various branches and sub-branches, each with its own proto-language. The phonological history of Proto-AN and its main descendants is known much better than that of Kadai or Miao-Yao, so we have quite reliable reconstructions of many PAN morphemes. However, the classification of the family remains quite uncertain. Many linguists accept different versions of Blust's provisional AN classification, which unfortunately is not a purely genetic one.[11] This uncertainty makes it very difficult to demonstrate that a reconstructed morpheme belongs to the proto-language level, rather than to some more recent one. There is no general agreement about how to solve the problem, but all Austronesianists agree that this is not a simple and straightforward task and that each etymology should be thoroughly investigated before it can be called Proto-AN (see, for example Mahdi's (1994) painstaking efforts with a few possible AN etymologies). There is however an extensive collection of AN etymologies published mainly by Dempwolff (1934-38) and Blust (1980; 1983-4, 1986; 1988; 1989), but more than half of all AN etymologies used in Benedict's comparisons are not found in these major sources. Instead he operates with his own pre-reconstructions based largely on the data Formosan languages. This a situation, plus widespread 'tuning' of proto-forms, makes the whole AN part of the JAT theory a collection of data which should be treated with great caution. A sample of AN pre-reconstruction can be *[C,s]ama 'green' based on a single form: Sediq sama 'green'. Sediq is an Atayalic language of Taiwan whose relationship with Proto-AN remains very obscure. An example of a 'tuned' proto-form is Benedict's *[q,ʔ]u(n)[z]ay instead of Demwolff's *ʔuḍay 'worm' (Benedict 1990: 263). The 'tuning' is needed to justify a comparison with Japanese uzi.

The history of Japanese can be understood only with the help of Old Japanese and Common Japanese-Ryukyuan reconstructions as most of the modern forms can be traced back to several different ancient forms. Intensive studies are undertaken in this area (see, for example, Martin 1987), but Benedict does not follow any particular reconstruction and uses modern Japanese forms or his own Old Japanese pre-reconstructions, which quite often are misleading (Vovin 1994).

Identification of comparisons

Given good quality initial data, the next step in checking a genetic claim is an analysis of comparisons included in it, and especially the evidence that these comparisons are real etymologies. The only way to demonstrate the genetic nature of comparisons is to analyse them with the help of systematic phonological correspondences between the languages under investigation and to

[11] Blust (1980: 11-12), for example, defines the Western-Malayo-Polynesian group not as a genetic unit with its own specific innovations, but rather as a residual group which did not undergo changes characteristic of the languages of other groups. The genetic nature of the primary split between Formosan and other languages is not properly motivated, and thus is also questionable.

apply them. Without them a claim remains a hypothesis, which can not be fully justified.

Often, however, a genetic claim is based only on a list of comparisons, not supported by systematic phonological correspondences or tools to eliminate loans. In such cases formal justification of genetic nature of comparisons is substituted by the intuition of the linguist. This immediately puts the claim beyond formal comparative evaluation and, strictly speaking, it should be rejected, as not of a comparative nature: one cannot argue against other people's intuition.

If two forms are included in a comparison, it means that the linguist who proposed this comparison believes that those forms represent two separate, independent and uninterrupted developments of a single proto-form. If this belief can be confirmed with the help of systematic phonological correspondence, and arguments that the comparison does not result from against borrowing, then we are dealing with an etymology. Otherwise, we have a comparison of unknown genetic origin. Cognates in an etymology can be similar to each other, or they can be quite different. For example in the Sino-Tibetan etymology 'eight': Chinese *ba*, Zangskar dialect of Tibetan *ɣat*, Burmese *šiʔ*, Luchuan *ʔhen*^{55c}. forms are quite different, while the meanings are identical. Transparent formal similarity between cognates is not however very important, when we are dealing with true etymologies: much more important is that it can be formally demonstrated that all such morphemes with different forms and meanings are various developments of a single proto-morpheme.

Working with languages which are not connected by a whole set of systematic phonological correspondences we do not have any formal means to prove that two morphemes should be included in a comparison. We can say only something like: 'Look, their forms and meanings are similar, so perhaps they can be traced back to a single source'. There is no way, however, to substantiate this suggestion. The danger of this situation, in the absence of any restrictions, is that we could bring together dissimilar forms, for example, Burmese and Yao morphemes 'fire': miː and tou^{4} (which are, in fact, not related) in a proto-form *toumi or *mitou and use this ghost as evidence in a genetic claim. To avoid such mistakes we rather deal with comparisons in which the resemblances reveal phonologically transparent connections. For each comparison we should be able to work out a correlation between the syllabic structures of resemblances and a correlation of individual phonemes in these structures. If on the basis of Proto-Zhuang-Tai *phram^{A} 'hair' and Proto-Kam-Sui *pram^{A} 'hair' a proto-form *p-ram 'hair' is suggested, I can not seriously argue against this pre-reconstruction, as it is based on clear similarity. But if this proto-form is connected with Proto AN *ra(m)but 'hairy' via two intermediate stages like

*p-ram < *[ra]p-ram[boc] = *[ts,tš]-r-a(m)boc > *ra(m)but

(Benedict 1990: 204-205) I have the right not to believe in it: too many changes need to be proposed to justify this comparison, and none of them have any supporting evidence.

Meanings of resemblances should also correlate rather simply. Ideally they should be synonyms in a broad sense. No unusual correlations are permitted at this stage of investigation and I cannot accept such distant semantic connection as 'above' / 'north', 'accustomed' / 'friend, companion', 'father or grandparents' / 'the god, thunder', as proposed in the first three comparisons given in the Benedict's work on JAT (1990: 161-162). Semantic relations like

'ant' / 'ant, 'back of a blade' / 'back, ridge', 'hind part' / back, behind' (cf. the next three comparisons in Benedict 1990: 162) are more convincing.

The restrictions placed upon comparisons do not mean that I a priori reject all non trivial etymologies. What I am saying here is, that at the stage of collecting data (the heuristic level of justification of a genetic claim) we should try to avoid any cases which are not straightforward, as they can lead to wrong results. Only after a genetic claim is proven and phonological correspondences are established, can we deal with the more obscure cases. At this stage (in proper comparative research) they should not affect our conclusions about genetic relatedness of the languages.

Many genetic claims found in the literature are of a heuristic type, with some of them intuitively more acceptable than the other. What makes the difference? In all generally accepted cases linguists are dealing with Young and Developed families where similarities between the languages are transparent, and anyone (whether speaker or linguist) can detect them. This allows scholars to compile lists of comparisons, but without comparative phonology they can not detect loans and separate them from etymologies.

In some cases, however, a simple procedure can help to do this. It is based on the following considerations. At least six major groups of morphemes can be distinguished in a language's lexicon. The first group - *descriptive morphemes* - includes morphemes which represent sounds, or activities accompanied by sounds. Such morphemes have a relatively high chance of being sound symbolic. Formal similarities based on onomatopoeia, idiophony and other types of sound symbolism, do not indicate genetic relationship. At the same time descriptive morphemes are not necessary sound symbolic. Quite often, however, it is difficult or even impossible to judge whether a descriptive morpheme is symbolic. In some cases, historical phonology can be of assistance; in others, the question remains open. Given the high probability of sound symbolism in such cases, it is preferable not to include descriptive morphemes in comparisons at the heuristic stage of investigation.

The second lexical group includes so-called *cultural* morphemes, or lexical morphemes with meanings related to various cultural ideas. As it is quite common for people to borrow ideas together with the appropriate words, we can expect a certain proportion of borrowings among the cultural morphemes of a language.

The third group includes morphemes which belong to the so-called *core lexicon*. The meanings of such morphemes are universal and are represented in most languages of the world, so it is less likely that such morphemes would be borrowed between languages. Of course, borrowings in the core lexicon are known, but the chances of runing across them here are usually not as high as for those in the cultural lexicon. Sound symbolic morphemes are also less common among core morphemes.

It is not simple to define a list of meanings which should be included in the core lexicon. Such meanings are represented, for example, in the 100-item and 200-item lists used in lexico-statistics, but a more extensive list could also be suggested.

The fourth group is formed by *grammatical morphemes*. In principle, these morphemes can be either original or borrowed, but normally we expect that grammatical morphemes are resistant to borrowing. There are also less chances for such morphemes to be of the sound symbolic type. From this point of view

grammatical morphemes are similar to core morphemes, but unlike the latter they are not universal and in some languages, like Classical Chinese, grammatical morphemes are extremely rare.

The fifth group is represented by lexical morphemes which can be called *environmental*. Their meanings are associated with various natural, floristic and faunistic phenomena: names of different species of vegetation, animals, birds and so on. The origins of such morphemes in a language reflects the history of its speech community. If migrations occurred, we expect to find many borrowings among these morphemes. In other cases they may remained unchanged.

The rest of the morphemes of a language belong to the sixth group. The origin of its members is hard to predict: they can be borrowed, of sound symbolic nature or be retained from previous stages of the language development.

The six groups identified here are not mutually exclusive and the origin of a particular morpheme cannot be predicted simply by its group membership. This membership, however, indicates its probable development: a morpheme from a cultural group will more naturally be borrowed than for a morpheme from the core lexicon. This observation is used in a technique for primary evaluation of a genetic claim. If languages are transparently related, they always have a certain number of comparisons among morphemes from the core lexicon. For Old and Macro families such comparisons should be found between forms of the proto-languages under consideration. There are no commonly accepted language families with no core comparisons, and if such comparisons are not found, a genetic claim seems unreasonable, even if it is supported by comparisons based on grammatical and other types of morphemes. This assumption leads to the following semi-formal procedure, suggested more than 20 years ago in some talks given by Jakhontov, and presented here in a modified form.

A check of a genetic claim can be based on the same lists of morphemes as those used for lexicostatistics. Each list includes the main, semantically unmarked translations of the 100 core meanings found in a particular variant (dialect) of one of the languages under investigation. Comparing lists by studying entries with the same meanings, a linguist identifies comparisons ,and separates them into original ones and loans. If the languages are transparently genetically related, they will always have a reasonable number of original comparisons. Without them a genetic claim is not valid.

Let us take as an example the relationship between three languages: Chinese, Tibetan and Burmese which belong to different branches of an Old Family, traditionally called Sino-Tibetan. Modern forms of these languages are so different that it is very difficult to detect similarities between Beijing's Mandarin, Lhasa Tibetan and Modern Spoken Burmese. An internal reconstruction of Chinese and evidence from the Tibetan and Burmese traditional orthography reduce the differences between the languages , and bring us to the level of a Developed Family (a situation similar to Indo-European with its archaic languages). At this level, similarities between the languages are more transparent, and a reasonable list can be collected. The main body of evidence that the three languages are genetically related is a list of several hundred comparisons (Peiros & Starostin 1966) which connect any pair or all three of these languages. They include lexical morphemes and pronouns, but comparisons of purely grammatical morphemes are not found. The number and

quality of comparisons rules out chance similarities as an explanation, leaving open only two possibilities: mass borrowing or genetic relationship.

Comparisons from the 100-item list indicate genetic relationship, as comparisons are found between any pair of these languages, as well as between all three of them:[12]

		Chinese	Tibetan	Burmese
1.	die	*sij?*	*ăchi*	*sij*
2.	ear	*nhə?*	*rna*	*na:*
3.	fire	*smə:j?*	*me*	*mi:*
4.	fish	*ŋha*	*ña*	*ŋa:*
5.	kill	*sra:t*	*gsod*	*sat*
6.	long	*draŋ*	*riŋ*	*hrañ*
7.	name	*mheŋ*	*miŋ*	*?a.-mañ*
8.	short	*to:n?*	*thuŋ-thuŋ*	*tui*
9.	sun	*nit*	*ñi-ma*	*nij*
10.	two	*nij-s*	*gñis*	*hnac*

		Tibetan	Burmese
1.	black	*nag*	*nak*
2.	bone	*rus*	*?a.rui:*
3.	dog	*khi*	*khuj:*
4.	eat	*za*	*ca:*
5.	eye	*mjig*	*mjak.ci*
6.	hand	*lag*	*lak*
7.	heavy	*l^ʒid*	*lij:*
8.	know	*šes*	*si.*
9.	liver	*mčin*	*?a.sañ:*
10.	meat	*ša*	*?a.-sa:*
11.	moon	*zla*	*la.*
12.	nail	*sen-mo*	*lak-sañ:*
13.	near	*thag-ñe*	*ni:*
14.	neck	*mdriŋ*	*lañ-paŋ:*
15.	nose	*sna-khug*	*hna-khauŋ:*
16.	not	*ma*	*ma.*
17.	road	*lam-kha*	*lam:*
18.	salt	*chwa*	*cha:*
19.	snake	*sbrul*	*mruj*
20.	star	*kar-ma*	*kraj*
21.	tongue	*lče*	*hlja*
22.	tooth	*so*	*swa:*
23.	tree	*šiŋ-sdoŋ*	*sac-paŋ*

		Chinese	Burmese
1.	dry	*ka:r*	*khrauk*
2.	horn	*kro:k*	*khjui*
3.	new	*sin*	*sac*
4.	night	*lia-s*	*ña.*
5.	sand	*sra:j*	*saj:*
6.	stone	*diak*	*kjauk*
7.	tail	*məj?*	*mri:*
8.	year	*nhi:n*	*hnac*
9.	yellow	*waŋ*	*wa*

[12] Starostin's Old Chinese reconstructions represent Mandarin words. Lhasa Tibetan and Standard Burmese forms are given in their traditional orthography. Most of the comparisons are well known (Shafer 1966; Benedict 1972; Peiros & Starostin ms).

		Chinese	Tibetan
1	I	*ŋha:j*	*ŋa*
2	louse	*srit*	*srig*
3	mouth	*kho:?*	*kha*
4	this	*te*	*ădi*
5	water	*tuj?*	*čhu*

Formal identification of the comparisons as etymologies is based on Sino-Tibetan comparative phonology as it is reconstructed in Peiros & Starostin 1996, but even without systematic correspondences the identity of the forms in most cases is quite obvious and is accepted by such linguists as Shafer, Benedict or Luce who worked without a complete set of phonological correspondences for these three languages.

Let us investigate what conclusions can be drawn from comparison of the Sino-Tibetan and Vietnamese 100-item lists.

In one comparison, the Vietnamese form is similar to those of all the other languages:

	Chinese	Tibetan	Burmese	Vietnamese
kill	*sra:t*	*gsod*	*sat*	*giết*

However, this comparison is not reliable. The Austroasiatic origin of the Vietnamese form is well known and the only formal similarity between the Vietnamese and those of other languages is the final *-t*.

No binary similarities between Tibetan and Vietnamese are found. Similarities between Burmese and Vietnamese, or Tibetan / Burmese and Vietnamese, are represented by one comparison each, remaining within the bounds of chance resemblance:

	Chinese	Tibetan	Burmese	Vietnamese
rain			*mui:*	*mưa*
tongue		*lče*	*hlja*	*lưỡi*

The majority of comparisons include a Chinese form:

	Chinese	Tibetan	Burmese	Vietnamese
1 fly	*pəj*	*(ăphir*	*pjam)*	*bay*
2 green	*che:ŋ*	*lʃ*aŋ-khu*		*xanh*
3 head	*s-lu?*			*đầu*
4 heart	*səm*			*trái tim*
5 lea	*lap*	*lo-ma*		*lá*
6 liver	*ka:n*			*gan*
7 near	*gən?*			*gần*
8 yellow	*waŋ*		*wa*	*vàng*

Even without any knowledge of the history of Southeast Asian languages we can suggest the only acceptable interpretation of these comparisons: they include chance similarities ('fly, 'leaf') and borrowings. As these comparisons are limited only to Chinese and Vietnamese and do not

include cases without Chinese, we can talk only about borrowings. The direction of borrowing (from Chinese to Vietnamese) is indicated by the fact that some Chinese forms ('green', 'yellow') in other Sino-Tibetan languages.

Let us now investigate a claim that Chinese is genetically related to Austronesian languages (Sagart 1994) using the same technique. Here one can find several types of comparison between the three Sino-Tibetan languages and Standard Malay. The first type is represented by Sino-Tibetan forms similar to Malay:

	Chinese	Tibetan	Burmese	Malay
1 die	*sij?*	*ăčhi*	*sij*	*mati*
2 dry	*ka:r*		*khrauk*	*kering*
3 long	*draŋ*	*riŋ*	*hrañ*	*pañjang*
4 road		*lam-kha*	*lam:*	*jalan*
5 sand	*sra:j*		*saj:*	*pasir*
6 tongue		*lče*	*hlja*	*lidah*

Most of these similarities are probably due to chance, but some may be loans. Comparisons which only hold between Tibetan and Malay are all of a chance nature:

		Tibetan	Malay
1	belly	*grod*	*perut*
2	sit	*bsdad*	*duduk*
3	stone	*rdo*	*batu*

No comparisons solely between Burmese and Malay are found. The comparisons with Chinese are more interesting:

		Chinese	Malay
1	cloud	*wən*	*awan*
2	egg	*ro:n?*	*telur*
3	foot	*kak*	*kaki*
4	hair	*pat*	*rambut*
5	root	*kə:r*	*akar*
6	salt	*lam*	*garam*
7	sleep	*duj*	*tidur*

Taken in isolation these could be treated as an indication of genetic relationship between Chinese and Malay. The addition of Burmese makes such a suggestion absolutely improbable. As it is clear that Chinese and Burmese are genetically related, one should expect to find comparisons between any these languages and Malay. The absence of reliable comparisons between Burmese and Malay leads us to interpret the data in exactly the same way as for the Sino-Tibetan languages and Vietnamese. The languages are unrelated, but there were some contacts between speakers of Chinese and Malay or of their ancestor languages.[13]

Now we can try to apply this procedure to the hypothesis that Austronesian languages are related to Kadai. Our Malay list reveals the similarities with the Siamese one:

[13] In fact the languages involved in the contacts were probably Proto-Chinese and a very ancient Austronesian language, possibly even Proto-Austronesian (Peiros & Starostin 1984, Peiros to appear).

		Malay	Siamese[14]	Proto-Kadai
1.	ashes	*abu*	*dau.B*	< **P-dau*^{B}
2.	black	*hitam*	*?dam.A*	< **?nam*^{T}
3.	die	*mati*	*ta:i.A*	< **I-ta(:)i*^{A}
4.	drink	*minum*	*?di:m.B*	
5.	dry	*kəriŋ*	*hə:ŋ.A*	
6.	eat	*makan*	*kin.A*	< **kɨVn*^{A}
7.	eye	*mata*	*ta.A*	< **I-nta*^{A}
8.	fire	*api*	*vai.A*	< **vVj*^{A}
9.	green	*hijaw*	*khieu.A*	< **R-mVɨ*^{A}
10.	know	*tahu*	*ru.C*	
11.	louse	*kutu*	*hau.A*	< **trau*^{A}
		(Malay	Siamese	Proto-Kadai)
12.	moon	*bulan*	*?dian.A.*	< **P-?nian*^{A}
13.	rain	*hujan*	*fon.A*	< **vaN*^{A}
14.	sand	*pasir*	*dra:i.A*	= Chinese borrowing
15.	this	*ini*	*ni.C*	
16.	tongue	*lidah*	*lin.C*	
17.	yellow	*kuniŋ*	*hliaŋ.A*	< **[C-]liaŋ*^{A}

This set seems more convincing than the previous one and indicates that Siamese and Malay are probably genetically related, as suggested by the Kadai-Austronesian theory (part of JAT). There is still a possibility that borrowings could account for the comparisons (see Thurgood 1994 who, however, does not operate with a Proto Kadai reconstruction), but I know of no well-supported Kadai-Austronesian comparisons from the cultural lexicon[15] and the possibility of loans predominantly entering the core lexicon seems to me rather strange.

A study of Yao, a Miao-Yao language which Benedict also includes in his JAT family gives, a different picture. Eight possible comparisons are found between Yao and Siamese with no specific comparisons between Yao and Malay:

		Yao	Siamese	Malay
1	bird	*no.8*	*nok*	(+ AN etymology)
2	die	*tai.6*	*ta:i.A*	*mati*
3	egg	*tc!au.5*	*khai.B*	
4	fish	*bjau.4*	*pla.A*	
5	long	*da:u.3*	*ja:u.A*	
6	salt	*dzau.3*	*klia.A*	
7	this	*na:i.3*	*ni.C*	*ini*
8	water	*wam.1*	*nam.C*	(+ AN etymology)

In two of these comparisons we also have resemblances from Malay. As no comparisons specific for Malay and Yao are known the data again can be

[14] The Siamese forms are given in transliteration. Proto-Kadai reconstructions are taken from Peiros, to appear.

[15] The whole list of comparisons which I can accept is included in Peiros to appear.

interpreted as an indication of contact between Siamese and Yao, but not as evidence for a direct genetic affiliation.[16]

Vietnamese and Khmer are genetically related: both belong to the Austroasiatic family. This fact is clearly confirmed by comparisons from their 100-item lists:

		Vietnamese	Khmer
1	bone	*xương*	*chəʔiŋ*
2	dog	*chó*	*chəkɛː*
3	earth	*đất*	*ṭiː*
4	foot	*chân*	*ʒaəŋ*
5	hair	*tóc*	*sok*
6	hand	*tay*	*ṭaj*
7	horn	*sừng*	*sənɛːŋ*
8	leaf	*lá*	*səlɨk*
9	louse	*chấy*	*caj*
10	meat	*thịt*	*sac*
11	neck	*cổ*	*kɔː*
12	new	*mới*	*thəmiː*
13	nose	*mủi*	*crəmuh*
14	one	*một*	*muəj*
16	root	*rễ*	*rɨk*
17	sand	*cát*	*khəsac*
18	sit	*ngồi*	*ʔəŋguj*
19	tail	*đuôi*	*kənduj*
20	this	*này*	*nəːh*
21	two	*hai*	*biːr*
22	water	*nước*	*dɨk*
23	what	*gì*	*səʔiː*
24	wind	*gió*	*khjal*
25	year	*năm*	*chənam*

If we now compare Vietnamese, Khmer and Yao lists, the results confirm a hypothesis of their genetic relation. Here we find triples and binary comparisons between any pair of languages:

[16] It is possible that the Miao-Yao and Austro-Tai families are related, but to prove it one should look for comparisons between reconstructed 100-item lists for the corresponding proto-languages.

		Vietnamese	Khmer	Yao
1	bone	*xương*	*chəʔiŋ*	*buŋ.3*
2	dog	*chó*	*chəkɛː*	*tcu.3*
3	horn	*sừng*	*sənɛːŋ*	*coŋ.1*
4	tail	*đuôi*	*kənduj*	*twei.3*
5	this	*này*	*nəːh*	*naːi.3*
6	two	*hai*	*biːr*	*i.1*
7	wind	*gió*	*khjal*	*dzjaːu.5*

		Vietnnamese	Yao
1	cloud	*mây*	*mou.6*
2	come	*đi*	*taːi.2*
3	eye	*mắt*	*mwei.6-tsiːŋ.1*
4	long	*dài*	*daːu.3*
5	round	*tròn*	*tɕ/un.2*
6	smoke	*khói*	*sjou.5*
7	you	*mày*	*mwei.2*

		Khmer	Yao
1	blood	*ʒhaːm*	*zjaːm.3*
2	rain	*bhliəŋ*	*bjuŋ.6*
3	tail	*kənduj*	*twei.3*

The same procedure can also be applied to the data Benedict uses to argue for a Japanese / Austro-Tai relationship. A comparison of the Proto-Japanese list with Austronesian (probably Proto-AN) reveals the following:

		Proto-Japanese[17]	Proto-AN
1.	drink	**nəm*	**ʔinum*
2.	eye	**màiN*	**mata*
3.	fire	**pə-i*	**Capuy*
4.	horn	**tùnwuá*	**t'uŋu*
5.	tooth	**pà*	**Cipən*
6.	tree	**kəi*	**kaSiw*
7.	who	**tá*	**t'a[y]i*
8.	yellow	**kúi*	**kuniŋ*

For of these Japanese forms (drink, eye, fire and tooth) have Altaic etymologies. Proto-Japanese reveals 25 comparisons with Korean and 15 with Tungusic (Starostin 1991:106) while neither Korean, nor Tungusic demonstrate any significant number of similarities with Proto Austronesian.

The result of applying this simple procedure to these languages suggests that we are dealing with four clear cut groups of them:

(1) Sino-Tibetan: Tibetan, Burmese and Chinese
(2) Austro-Tai: Siamese and Malay
(3) Vietnamese, Khmer and Yao
(4) Japanese (with other Altaic languages).

[17] Proto-Japanese forms and their Altaic etymologies are taken from Starostin 1991.

These conclusions, however, are preliminary and can be accepted only at the heuristic level of argumentation. It is worth remembering that the procedure cannot detect that groups (2) and (3) are possibly related, nor suggest any classification of languages within these groups.

The following considerations are important if the procedure is used:

1. Any two transparently related languages always show a certain number of comparisons from the 100-item list, usually more than 12-15. About 5 comparisons will usually be found between any two languages due simply to chance factors, and they do not indicate a genetic relationship. A lack of comparisons means that the languages cannot be directly connected to each other. To prove that they are remotely related one needs to study their proto-languages looking for comparisons between their reconstructed 100-items lists.

2. Any genetic claim based on 100-item lists should include comparisons from at least three languages: a binary comparison can leads to distorted results, as for the Chinese-Vietnamese relationship. A genetic claim not supported by a system of phonological correspondences should be based on interpretation of data from several languages, to aid in the detection of possible loans and other perturbations.

3. Comparative study of languages requires their systematic investigation: a comparable amount of data should be used and presented for each language involved. We could not take seriously a claim that languages A, B and C are genetically related which is based on twenty comparisons between languages A and B and on another twenty found in A and C, but not in B.

Phonological correspondences

Related languages always have comparisons involving forms from their core lexicon, usually supported by comparisons from other lexical groups. The presence of these comparisons, however, is not in itself enough to prove a genetic claim. Proof is possible only where a set of systematic phonological correspondences is presented. Without them, such a claim remains a more or less plausible hypothesis.

Phonological correspondences established from the whole collection of comparisons would be of two types:

(i) those connecting phonemes of common origin, and

(ii) those connecting phonemes in forms which are not genetically related (usually a result of borrowing).

A phonological correspondence which brings together reflexes of a particular proto phoneme or other features of a proto-language is called a *systematic correspondence*. Simply looking at a correspondence, however, we can never say if it is systematic: a reconstruction of the entire phonological system of a proto-language is needed before a correspondence can be identified with any certainty as being systematic.

A phonological correspondence supported in a sufficient number of comparisons is called a *regular correspondence*. The expression 'sufficient number of comparisons' is rather vague and usually depends on the number of comparisons found. If we have, for example, a thousand comparisons, a phonological correspondence supported by a hundred of them is regular, while a correspondence supported by only two comparisons is not. Quite often it is very difficult to decide if a correspondence is regular. What if a correspondence was based on two or three comparisons out of total a hundred reliable comparisons?

Regular correspondences can be found in comparisons which connect morphemes of common origin as well as borrowings. In cases of the latter type, such as between Chinese, Vietnamese and Japanese, regular correspondences occur only in situations of mass borrowings. Despite the fact that a system of regular phonological correspondences is known, these languages are not genetically related. Vietnamese and Japanese have intensively borrowed from Chinese during a rather limited period of time and the regularity of phonological correspondences reflects this fact.

What is necessary for proof of genetic relatedness, then, is a set of systematic (though not necessary regular) phonological correspondences. This set connects the phonological systems of the languages under investigation, which means that for each phoneme a of language A we have to find corresponding phonemes (sometimes Ø) in all other languages. Usually, most of these systematic correspondences will be regular, supported by sufficient number of examples. A systematic correspondence can, however, be associated with a rare feature of the proto language and for this reason may be represented only in few etymologies. It is very important that the proposed set of systematic correspondences should connect all elements of forms included in a comparison, rather than being correct, say, only for initial consonant or tones.

Practically speaking, for any genetic claim we need to have tables of systematic phonological correspondences between all the languages discussed in the claim. Such tables should be given for all parts of their phonological systems, including syllabic structures, consonants (initial, medial, final as well as consonantal clusters), vowels and, if necessary, such suprasegmental elements as tones, registers or stress patterns. With the help of such tables we should be able to check whether the grouping of particular morphemes into etymologies is convincing or not. The regular correspondences in these tables should be identified and we can expect that they will be found in most etymologies.

For the genetic claims mentioned above, systematic phonological correspondences are given only for the SC theory. Here they connect only syllabic structures and consonants. Unfortunately, the published correspondences do not connect other parts of the phonological systems, primarily vowels, of the proto languages compared. Application of the correspondences to the forms included in comparisons shows that they are fairly consistent, and do not contradict each other. This means, at least for me, that it is highly probable that the Sino-Caucasian theory is correct, but the whole set of systematic correspondences is needed to provide the body of evidence formally required for the proof of the claim.

Systematic correspondences are not known for Austro-Tai and Miao-Austroasiatic hypotheses, which are supported by limited numbers of comparisons. As those comparisons can hardly be explained through borrowing, it is likely that further research would lead to the discovery of systematic correspondences among the proto-languages constituting each of these two families. The Sino-Austronesian and Japanese / Austro-Tai hypotheses are not supported by convincing comparisons, it is not surprising that sets of systematic phonological correspondences are not found. This means that both hypotheses should be rejected.

References

Benedict, Paul. K.
1942 Thai, Kadai and Indonesian: a new alignment in South Eastern Asia. - *American Anthropologist,* n. s. 44 (4): 576 - 601
1972 *Sino - Tibetan: A conspectus.* - Cambridge: Cambridge University Press.
1975 *Austro - Thai. Language and culture*. New Haven: Human Relations Areas Files Press.
1990 *Japanese / Austro-Tai* (Linguistica Extranea Studia 20). Ann Arbor: Karoma Publishers.

Blust, Robert A.
1980 Austronesian etymologies. - *Oceanic Linguistics*, 19/1-2: 19 - 181
1983-84 Austronesian etymologies II . - *Oceanic Linguistics*, 22/23: 29 -14
1986 Austronesian etymologies III. - *Oceanic Linguistics*, 25/1-2 : 1 -123
1988 The Austronesian homeland: a linguistic perspective. *Asian Perspectives* 26: 45 -67.
1989 Austronesian etymologies IV. - *Oceanic Linguistics*, vol. 28: 111 - 180

Dempwolff, Otto
1934-38 *Vergleichende Lautlehre des Austronesischen Wortschatzes*. Berlin: Dietrich Reimer 3 vol.

Jakhontov, Sergej E.
1980 Ocenka stepeni blizosti rodstvennyx jazykov. In V. N. Jarceva (ed.) *Teoreticheskie osnovy klassifikacii jazykov mira*. 148-157. Moscow: Nauka.
1981 Sootvetstija finalej v dialektax mjao.*Vostokovedenie*, vol. 8: 63-72.

Li, Fangkuei
1977 *A handbook of comparative Tai.* (Oceanic linguistics Special Publication 15) Honolulu: University Press of Hawaii.

Lombard, Sylvia J.
1968 *Yao-English dictionary* (Data paper 69, Linguistics series 2). Ithaca: Cornell University Southeast Asia Program.

Mahdi, Waruno
1994 Some Austronesian maverick protoforms with culture-historical implications. *Oceanic Linguistics* 33 1: 167-230; 2: 431-90.

Martin, Samuel E
1987 *Japanese language through time*. New Haven: Yale University Press.

Matisoff, James A.
1988 Proto-Hlai initials and tones: a first approximation. In J. A. Edmonson, D. B. Solnit *Comparative Kadai: linguistic studies beyond Tai.*: 289 - 321. Summer I nstitute of Linguistics and The University of Texas at Arlington Publications in Linguistics 86.

Nikolaev, Sergej L and Sergej A. Starostin
1994 *A North Caucasian etymological dictionary*. Moscow: Asterisk Publishers

Peiros, Ilia I
to appear *Linguistic Prehistory in Southeast Asia.* Pacific Linguistics C-

Peiros, Ilia I. and Sergej A. Starostin
1984 Sino-Tibetan and Austro-Tai. *Computational analysis of Asian and African languages*. vol.22: 123-127.
1996 A Copmparative Vocabulary of Five Sino-Tibetan Languages. Fascicles 1-6. The University of Melbourne: Department of Linguistics.

Purnell, Herbert C Jr.
1970 *Toward a reconstruction of Proto-Miao-Yao*. PhD dissertation, Cornell University. Ann Arbor: University Microfilms International.

Sagart, Laurent
1993 Chinese and Austronesian: evidence for genetic relationship. Journal of Chinese Linguistics 21: 1- 62.
1994 Proto-Austronesian and Old Chinese evidence for Sino-Austronesian. Oceanic Linguistics 33/2: 271-308.

Shafer, Robert
1976 *Introduction to Sino-Tibetan*. Vol. 1. Wiesbaden: Otto Harrassowitz..

Starostin, Sergej A.
1982 Praenisejskaja rekonstrukcija i vneshnie svjazi enisejskix jazykov. In E. A. Alekseenko (ed.) *Ketskij Sbornik*. 144-237. Moscow: Nauka.
1984 Gipoteza o geneticheskix svjazjax sinotibetskix jazykov s enisejskimi i severnokavkazskimi jazykami. *Linguisticheskaja rekonstrukcija i drevnejshaja istorija vostoka*. P.4: 19-38. Moscow: Nauka.