The Vietnamese etymological dictionary and 'new' language families

(Project in progress)

Ilia Pejros
Department of Linguistics
The University of Melbourne, Australia

The Austroasiatic language family which includes Vietnamese has been the subject of comparative investigation since the last century. However, it is only now, at the end of the twentieth century, that linguists are able to undertake truly detailed comparative study of the family. From this point of view, Austroasiatic is a typical 'new' language family, in contrast to the 'old' families with long and successful comparative traditions: Indo-European, Semitic and many others. The label 'new' does not indicate that the family is young, and formed only by closely related languages, but rather provides us with information about the level at which comparative studies in the family's linguistic prehistory have been conducted.

The following features are common to 'new' families:

• Lack of detailed synchronic descriptions for the majority of languages included in the family. There are usually good and reliable descriptions (phonological studies, grammars and dictionaries) for just a few of the major languages of the family, with the rest known only superficially. In the Austroasiatic family less than 15 languages are known well enough to be included in comparative study.

• No generally accepted genetic classification. In the Austroasiatic case, linguists are able to identify primary groups of closely related languages, like Vietmuong or Munda, but it remains unclear how these groups are related to each other.

• Reconstructions are available for some of the primary groups of the family. Only three reliable reconstructions of Austroasiatic groups are available to

me: Vietmuong (Sokolovskaja MS), Monic (Diffloth 1984) and Katuic (Pejros 1996).

• There is no convincing reconstruction of the proto-language of the family, and linguists usually have to deal with hypothetical structures which are not based on thorough comparative investigation.

• A lack of reconstructions means that it is difficult (and sometimes even impossible) to identify forms borrowed from one related language into another.

One of the basic principles of modern comparative linguistics is the aspiration towards completeness which applies both to the data and its interpretation. The comparative method requires that a reconstruction should be based on a thorough study of all the languages included in the family, and a detailed comparison of each pair of them. 'Mass' comparison (in which a word from language A is compared with a word from language B, while another word of A is compared with a word from language C without any attempt to find its counterpart in language B) is not a valid procedure. All comparisons should be conducted between pairs of languages (A and B, B and C, A and C, etc.), and the overall investigation should include a thorough comparison for each of the possible pairings of the chosen languages. The reasons for this requirement are clear: the aim of a proper comparative study is to find regular patterns which connect the systems of all languages under investigation, in order to reconstruct the ancestral proto-language. This necessitates a knowledge of relationships between the systems of all the languages studied. Only a comparison of each and every pair of these languages, one by one, is sure to discover all possible connections between their systems.

Completeness is thus a central principle of modern comparative studies. Applying this principle to the study of 'new' families, the first stage of our investigation must involve:

(i) compilation of a provisional comparative dictionary, which includes morphemes of presumed common origin found in any two, three or more languages of the family;

(ii) establishment of a set of phonological correspondences which connect all the phonemes of all languages of the family.

These two components lead linguists to:

(iii) identification of systematic phonological correspondences which are traces of phonological distinctions of the proto-language
(iv) reconstruction of the system of proto phonemes; and
(v) reconstruction of lexical and grammatical morphemes of the proto-languages (with the aid of the phonological reconstruction and the provisional comparative dictionary). A collection of these reconstructed proto-forms is usually called an 'etymological dictionary' of the family. An etymological dictionary differs considerably from the provisional comparative dictionary used in its completion, as it comprises only genetically related morphemes of the languages, omitting resemblances which are due to borrowing, chance similarities, local influences or other non-genetic factors.

It is clear, however, that everything in comparative linguistics depends on success in the completion of a provisional comparative dictionary, and on our ability to interpret its evidence. This is the basis upon which a linguist can conduct a morphological reconstruction, create a genetic classification of the family, and engage in other comparative research.

An etymological dictionary for a particular language – a culmination of comparative investigation into that language's history – is based on intensive lexical studies and on a deep knowledge of the historical phonology of the language and its family. Does this imply that any attempt to compile an etymological dictionary for a language of a 'new' family is *a priori* premature? I do not think so and for several years now I have been completing an etymological dictionary of Vietnamese. The theoretical foundations of this project are discussed below.

As mentioned above, a 'new' family is usually known only partially with detailed reliable descriptions being unavailable for most of its languages. In dealing with such a family, then, we need a strategy which will meet the principle of completeness. One strategy is to include data from every language or dialect mentioned in the literature, regardless of the possibility that this data is by no means complete. If, for example, a language is known only by a list of 200 words, we could include it in our investigation, but we would not expect to be able to draw detailed conclusions about this language. If the number of such poorly recorded languages is significant, then a provisional comparative dictionary and

phonological correspondences based this strategy will be complete, but not adequate for a reconstruction.

Another possibility is to concentrate only on well-known languages but extract all possible information from their comparison. In such a strategy the emphasis is on the complete interpretation of the data rather than on the completeness of the data itself. Both of these strategies are represented in the literature, but I think that the second one is more appropriate in the circumstances of 'new' language families such as Austroasiatic. My project is therefore based on a study of several languages chosen as primary sources, and an attempt is made to identify all possible comparisons between these languages and Vietnamese.

The starting point of the project is a list of common Vietnamese morphemes represented in the major modern Vietnamese dictionaries. To it I have added some archaic morphemes whith interesting etymologies. Recent borrowings, mostly from European languages, are not included. Altogether the list comprises about 5,000 entries.

Vietnamese, together with many closely related languages, belongs to the Vietmuong group of Austroasiatic. In principle, this group should be investigated before a comparison with other Austroasiatic languages begins. However, it is not possible to compile a full scale Vietmuong comparative dictionary because published data is not available for most Vietmuong languages (Barker 1993). With the exception of Vietnamese, only one Muong dialect has been described in any detail (Materialy 1987), although there are short dictionaries of Ruc (Nguyen et al. 1988) and Thavung (Ferlus 1979). Nonetheless the history of the group has been the subject of intensive investigation, and numerous reconstructions have already been proposed (Barker 1963; Barker 1970; Ferlus 1975; Thompson 1976; etc.). In my opinion the best reconstruction is Sokolovskaja's (MS), which is based on interesting and important data including Nguyen Van Tai's dissertation, and unpublished field materials of joint Russian-Vietnamese linguistic expeditions. Sokolovskaja identifies more than 700 cognates showing good semantic and phonological correspondences between the selected languages. The Proto-Vietmuong reconstruction includes many initial consonantal clusters, and some disyllabic roots. There is no evidence of tonal oppositions, and the tones of modern languages are explained as having developed from two suffixes *-*x* and