FAMILY EVOLUTION, LANGUAGE HISTORY AND GENETIC CLASSIFICATION[1]

Ilia Peiros

Lexicostatistics[2] remains a pariah of comparative linguistics. Sometimes linguists agree to use it as very preliminary tool of investigation, but more often they totally reject it on the basis of general observations. No present day discussion of lexicostatistics has been conducted so far.

When discussing lexicostatistics, we need to address the following issues:

- What are the methods used in lexicostatistics?
- Why is it used?
- How can it be used?
- How does it fit into the broader framework of comparative linguistics?
- How does a lexicostatistical classification of a family correspond to other classifications of the same family proposed used in comparative linguistics?

The theory of lexicostatistics is the subject of a monograph entitled 'Lexicostatistics revisited' which Starostin and I are working on. Currently the structure of the monograph is the following:

Chapter I. The theory of lexicostatistics.
Chapter II. Glottochronology.
Chapter III. Lexicostatistics as a heuristic.
Chapter IV. Case Studies.

Here I present some parts from Chapter I, as they are written by me in Melbourne in 1997 on the basis of discussions with Starostin which started about twenty years ago and resume every time we see each other. Unfortunately, the distance between Melbourne and Moscow (I hope only a geographical one!) did not allow me to finalise this text with Starostin and to present it as a joint publication.

[1] I am grateful to Marc Durie for his help and inspiration. The final stage of this project has been supported by an Arts Faculty Completion Grant of the University of Melbourne.

[2] We distinguish here lexicostatistics (a method of genetic classification of languages) and glottochronology (a method of obtaining absolute datings in comparative linguistics).

Therefore I am publishing it under my name, hoping, however, that the article correctly reflects the essence of the future monograph.

§1. Any human community can be represented as an informational network which facilitates direct or indirect communication between its members. Such a reliable permanent informational network connecting all its members is probably the most important condition of the very existence of a community: if communication is not maintained, a community has little chance to survive.

The media of communication is always human languages which form a linguistic repertoire of a community (a set of languages used in it). This repertoire is not necessarily limited to one language, and multilingual communities are not less typical than monolingual, but the necessity of reliable informational exchange is normally based on the fact of shared language knowledge[3].

This leads us to the well-known observation that the main function of any natural language is to maintain informational exchange between people who belong to the same community.

Each community has certain views about languages used by its members and in other communities. These linguistic views are not necessarily identical to the views accepted by professional linguists. Therefore we will distinguish between the two notions, 'language' and 'sociolanguage'. Although two speakers of the same language can use it with sometimes quite noticeable differences, when discussing common topics they will always understand each other. This makes the criterion of mutual intelligibility to be essential for the notion 'language'[4]. Two speech varieties belong to 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:

One language - one sociolanguage

Speakers of Hungarian know that they use the same (socio)language. Minor differences do not prevent mutual intelligibility, which means that they belong to the same language.

One language - two or more sociolanguages

[3] This is not true for more complex entities of human organisation, for example empires, like British or Roman.

[4] Dialectal chains, if they exist, do not contradict this claim (Peiros 1989)

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 an excellent example of this type of relation. It is well known that the differences between some of them are not less than between major Slavonic languages (where independent language status is generally accepted) and mutual intelligibility is normally not possible. At the same time speakers of Chinese 'dialects' know that they speak the same (socio)language and are prepared to defend this claim.

In comparative linguistics the main focus is on languages, not sociolanguages. Therefore a language family is seen as formed only by languages, where differences can be evaluated by formal methods without an appeal to the views of their speakers.

§2. Any traditional community lives in its habitual world with a well-known environment, usual activities, customary social relations and other predictable features. In such a life, similar events usually occur more or less frequently and it is always known how a member of the community is supposed to act, what should be said and what kind of response is expected from other people involved. This knowledge and the ability of predictions are essential for the well being of any community.

Using linguistic tools, members of a community are able to convey any type of information regarding their everyday activities, typical situations, environment and so on. It is still to be investigated how complete and precise is this information, but from the theoretical point of view it is clear, that all basic informational demands of a stable traditional community are met by language(s) used in it. It does not mean, however, that any type of information can be easily communicated in any language, as often speakers have trouble when trying to express alien ideas in their own language. In any typical situation, however, each community has sufficient linguistic means to communicate appropriate information. This information is always community-specific and it is practically impossible, for example, to use Russian when talking about trees of the Indian jungles or to discuss ancestor rituals of Hmong in English.

The predictable way of life can be maintained while the community lives in the world with no significant changes: it occupies the same territory, it is not forced to change its activities, its neighbours and cultures also remain the same. A migration to a new environment or contacts with new cultures would undermine stability and the community would face the problem of adaptation to previously unknown situations. Such an adaptation always affects the existing linguistic repertoire. The community can either adopt a new language which is better prepared for the new life, or adjust the old one to new demands.

§3. Changes in the life-style are usually related to adaptation of new cultural ideas: new objects of material culture, new skills and views and if the community migrated, knowledge of the new environment. Over a certain period of time these changes would be reflected in the community's language(s).

It is generally believed that the culture of a community is represented in its language, mainly in its cultural lexicon, which is formed by words related to various cultural ideas. New ideas are usually represented either by old words with modified meanings, or by borrowings, as it is quite common for people to borrow ideas together with the appropriate labels (words) (see, for example, Simpson 1985). Therefore, the cultural lexicon of a language

- is historically not very stable and can be significantly changed over a short period of time;

- can include many borrowings reflecting the process of cultural adaptation.

A migration to a new territory with unknown vegetation and animals can also cause significant changes in the so-called 'environmental' lexicon of the language, which is formed by various words with meanings related to the natural world. Under certain circumstances we can expect to find here quite significant changes reflecting the differences of the two territories. In some cases the reorganization of 'environmental' lexicon is achieved mainly by changing meanings of original words and forming new complex expressions (see, for example, Biggs 1991). In other cases the main emphasis is put on loans.

There is, however, a particular part of a lexicon, which is less affected by changes in the community's life. We can identify more or less a universal set of ideas known in all or nearly all human communities regardless of their level of cultural development, territories occupied and other properties. These ideas are quite basic and simple: 'moon', 'sun', 'man', 'water', and many others. No doubt that languages differ considerably in the ways they present these ideas, but for any language, there are always means to represent them, unlike environmental or cultural ideas. Therefore it seems useful to talk about 'core information' and **core meanings** representing it. These core meanings can be represented in languages in many different ways, but it seems to be possible to identify such meanings across languages. As there are no obvious or universal reasons why core meanings should be changed in time, they are generally more stable than, say, environmental or cultural meanings and normally are better preserved in languages.

§4. Any human community *N* always has at least one language *T* known to its members. This language can be either inherited from the previous generations of speakers or be learned through contacts with other people. There are no recorded cases of glottogenesis[5] - the creation of a new language in total isolation from other languages and apart from the earliest periods of human history (not studied by comparative linguistics), glottogenesis is not possible.

The relation between community *N* and its language *T* always reflects the current status of *N*: whether it is stable or is in the process of formation or disintegration.

Over the whole period of stability, any two consecutive generations of *N* share the same language *T* learned from their parents and are able to communicate using it. No other languages are acquired by *N* in this period and its linguistic repertoire remains unchanged. The norms of language usage accepted by the majority of the community's members govern communication and sanction all changes to the system of *T*.

The period of formation of a new community is actually the period of establishing a new informational network to connect all its potential members. Among the important considerations here are:

- the necessity to have a reliable network functioning with minimum distortions;

- the necessity to have this network operational in the shortest possible time. (The longer the period, the lesser the chances are for bringing together potential members),

[5] Formations of pidgins and creoles always take place in the situation of linguistic contact.

- the necessity to have a user-friendly network, especially suitable for potential members in more prominent social positions.

All these considerations indicate that the only acceptable strategy in creating such a network is an adoption (at least as the basis for communication) of a language, already used for such purposes.

A new community can be formed either through disintegration of an earlier community or through crystallization from previously distinctive human groups.

It is important, however, to remember that in both options, languages are always either inherited or borrowed.

The process of a community's disintegration is related primarily to the collapse of a single system of norms accepted by all its members of the original community. Due to various extra-linguistic reasons (migration, political turmoils, etc.), norms adopted in community *N* lose their authority over potential members leading towards disintegration of *N* into several daughter groups. At very early stages of disintegration, norms governing the behavior of members are more or less identical across these groups, reflecting their common origin. Later, each group begins to develop its own norms no longer following a common pattern. This can be caused by difficulties in maintaining communication (the groups are not in contact any more), or is the result of purposeful attempts to create and maintain a new identity ('We should not dress like them', 'We do not say this word', etc.). Over a certain period of time, the accumulated effect of changes caused by the distinctive sets of norms would wipe away most of the features of common origin.

When a new community is crystallized from previously different groups it either adopts a language used by one of these groups, or borrows one from neighbours. Both options are explained by the following model. In any stable human group one can always identify a dominant subgroup which is also the center of the informational exchange. The language of this dominant group will inevitably be used for communication. If another language is chosen, it would mean the loss of influence from a previously dominant group (it would be in a disadvantaged position in communication) and the rise of another dominant group (associated with the chosen language). The dominant group can be a part of a new community, or it can exist separately from it, as it happens in the formations of creoles. In both cases, however, a language, which is tightly associated with power, is adopted as the basis of communication.

Therefore a crystallizing community acquires its language:

(i) either by inheriting the language of its ancestors (of the whole new community) or some of them (of a certain part of it); or

(ii) through borrowing from another community.

Under no conditions could glottogenesis take place.

From the above discussion it follows that we need to distinguish between the history of language L (changes in its system: phonology, morphology, lexicon, and so on) and the linguistic history of a community: its maintenance or language shift[6]. There are three main options in the linguistic history of community *N* with language *T*:

(i) *N* has inherited *T* from previous stages of its development;

(ii) *N* has borrowed *T* from another community *T*, either adding it to its repertoire, or using *T* instead of its original language (a shift to *T*);

(iii) *N* has stopped using *T* for any type of communication.

In the cases of (i) and of (ii) *T* would possibly undergo many significant changes and thus it would be rather different from the original one, but under no circumstances would a new language (not based on a language or languages, which already exist) be invented by *N*.

An uninterrupted development of *T* is related to options (i) and (ii), reflecting two possibilities of language acquisition by its speakers: through inheritance or borrowing.

§5. Every spoken language is the subject of a permanent process of changes, based on the high redundancy of human languages.

Change φ <α, β, Q, t> is a process which occurs in the system of a language and can affect any element of it:

α is the initial stage of change φ;

β is the outcome of change φ;

Q is a set of conditions under which change φ took place;

[6] If in the process of community development language A is supplanted by a borrowed language B, we will talk about a shift from A to B.

t is a period of time when change φ took place.

Depending on the relation between α and β, we can distinguish:

	Period A		Period B
drifts	α	=>	β
losses	α	=>	ø
additions	ø	=>	β

In a drift, β is an interrupted development of α and is its reflex.

A change can be caused externally or internally. An external causation of a change can be due to the influence of another language, or it can reflect a conversion of several languages within a linguistic area. However, in real practice it is sometimes difficult or even impossible to determine the true nature of the causation.

Changes in a language can either be triggered or free. If a change is caused by another change, which happened earlier, we are dealing with a triggered change. Otherwise a change is a 'free' one. Extralinguistic features, which are often behind the linguistic changes (especially in the lexicon), are not seen here as triggers. Therefore, if a new word is created to represent a new idea, this new idea is not seen as a triggered change. But a split of vowels which took place as a result of development of register is seen as a triggered change.

Altogether we can identify 12 different types of changes:

	Internal		External	
	Triggered	Free	Triggered	Free
drifts	1	2	7	8
additions	3	4	9	10
losses	5	6	11	12

1. Triggered internal drift
   For example: the change of meaning of the English word 'hound' in the process of the adoption of the word 'dog';

2. Free internal drift
   For example: the retention of Old English words in their modified modern forms;

3. Triggered internal additions
   For example: the development of the fixed word order caused by losses of various morphological distinctions;

4. Free internal additions
   For example: the creation of new compounds to represent new ideas;

5. Triggered internal loss
   For example: the loss of a distinction between two noun cases, caused by the loss of final vowels;

6. Free internal loss
   For example: the loss of words - labels of artefacts not used anymore;

7. Triggered external drift
   For example: the development of articles under the influence of another language;

8. Free external drift
   For example: the usage 'Ja ne dumaju' (literally 'I don't think') instead of 'Ja v etom ne uveren' to map the English expression 'I don't think so' by Russian migrants in the English speaking world;

9. Triggered external addition
   For example: the development of classifiers in many Southeast Asian languages (the actual forms can develop due to internal drift)

10. Free external addition
    For example: a lexical borrowing as a label for a new concept;

11. Triggered external loss
    For example: a loss of first syllables in many Southeast Asian languages (e.g., Vietnamese or Chamic) as a result of regional convergence;

12. Free external loss
    For example: a displacement, when an original word is lost and its functions are now performed by a borrowing.

These types of changes can be found in the history of any language whose development is, in fact, the process of accumulations of changes' outcomes.

§6. Studying the history of language L we need to identify at least:

(i) individual changes, which took place in the process of L formation and afterwards. This study always includes a description of the four components (<α, β, Q, t>) of these changes;

(ii) pairs of triggers and triggered changes;

(iii) relative chronology of the changes.

In such a study, we are supposed to investigate both external and internal caused changes and if possible, to specify the sources of their causations. In many cases, however, this cannot be done. If, for example, the language which was a source of intensive borrowings for L is not known, we often cannot identify the borrowed words in L. This, however, would not prevent us from describing the history of L.

Every single language has its own linguistic history which is always different from the history of any other language, because the changes and their chronology are always language specific.

The accumulation of various changes in a language's system is the process of language development. Changes affect all elements of a language system with no exceptions: nothing in a language is immune to change.

Over a certain period of time the language used by descendants of group *N* could become quite different form language *T* once spoken by *N*, even without a language shift. Accumulated substitutions can wipe away original features of *T*, leaving us with a question of how we can demonstrate that language L is a development of *T* and not, say, ***R***.

This leads us to the notion of **language continuity.** We will talk about language continuity from period P to period P', if for the whole time t elapsed between these periods, for every consecutive pair of generations of speakers, the core information was conveyed mainly with the help of linguistic expressive means of the same origin. In other words, language L is a continuation of *T* if its expressive means for core meanings have been inherited from *T* and are mainly the result of various drifts, rather than additions.

What does it mean, however, 'mainly with the help of linguistic means of the same origin'? The formal answer would be that if 51% of such meanings in language A' came from language A and 49% came from B, we talk about language continuity from A to A'. If later, the balance would change and some more linguistic meanings from B would substitute meanings from A, we would have to assume that a language shift took place and A' is now a continuation of B, rather than A[7]. It does not mean, however, that we accept the idea that a genetic affiliation of a language can be changed in time. We have only registered a language shift: before the change of that balance, people spoke a language which was a continuation of A which was full of borrowings from B, while after the shift they began to use a continuation of B with borrowings from A.

§7. Language *T* is often subject to the process of disintegration accompanied by the formation of two or more new languages, each being a language continuity of *T*. The following model explains this mechanism:

> In a certain period, there was a community *T* associated with language *T*. The usage of *T* was governed by an extremely complex set of norms which were more or less obligatory for all members of the community. The development of *T* was also governed by these norms, and only the changes approved by these norms were incorporated into the language system.
>
> Due to various extra-linguistic reasons, community *T* began to disintegrate into several groups which supplanted *T*. At the very early stages of disintegration, the norms of language usage were more or less identical for all these groups, reflecting their common origin. Later, the groups began to adopt norms, which were not necessarily shared by all of them. These different norms sanctioned different changes to the original identical language system. Their accumulated results caused a split of a previously common language into its daughter-languages L, L', L", each having a language continuity from *T*. Over time, these new languages began in their turn to disintegrate following the same model as their ancestors. This led to a formation of a language family which includes all language continuities from *T*.

[7] It is worthwhile to mention that no recorded cases of such situations are known to us.

It follows from the previous discussion that if L is a result of development of *T*, under no circumstances would it become a development of *T'* which is not a continuation of *T*. Obviously, hypotheses about genetic affiliation of L can be changed, but not the affiliation itself. A community can also change its language, and it is highly possible that its descendants would use a language of another genetic affiliation. In such cases we are dealing, however, with a language shift, rather than with a shift of genetic affiliation: a position of a language in a language family is permanent and does not change in time.

Let us introduce some more technical terms:

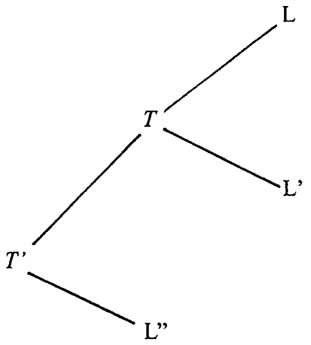
If L is a language continuation of *T* we will call L a 'descendant' of language *T*, while *T* is the ancestor of L.

Language L is a **daughter-language** of *T* if:

(i) there is language continuity from *T* to L; and

(ii) there is no such *T*' which is an ancestor of L and a daughter-language of *T* . In Figure I all five languages are in the family of *T*, but only four of them are related directly: *T*/L, *T*/L', *T'*/L" and *T'*/*T*.

Figure I



All languages with a common ancestor form a **language family.** Languages which belong to the same language family are genetically related. In other words, if two languages L and L' are genetically related they always have a common ancestor *T* and they are uninterrupted continuations of *T* in time. Strictly speaking, therefore, if we want to demonstrate that the languages are related, we should be able to present their common ancestor.

Two languages are specifically related if they are daughter-languages of the same ancestor language. L and L' in Figure I are specifically related, as are *T* and L". Languages L and L" are not specifically related as they have different direct ancestors. Specifically related languages, let us call them **sister-languages,** form separated groups within the family, sometimes called 'branches' of the family.

§8. When comparing languages *T* and L which are separated by a significant period of time, one needs to address two logically independent questions:

- is there language continuity between *T* and L?

- is feature β found in L a reflex of feature ***a*** found in *T*?

Answers to these questions are not necessarily interrelated and often, despite language continuity between *T* and L, β is not a development of ***a*** as it is for example a borrowing from an unrelated language. Usually we do not have much trouble interpretating such situations, and we simply say that *T* and L are genetically related but that J is a borrowing. However, in the literature, one can find discussions of more complex situations of mixed languages, namely languages which have had more or less equal amounts of elements of different origin, such as say, in L, where the source of its lexicon is language A and the source of its morphology is language B. The question typically asked for such situations is: how can we identify the genetic affiliation of this language - as a development of A or of B? According to the definition given above, the language continuity is determined by linguistic means associated with core meanings, and so the language which is the main source of such meanings in L should be recognized as its ancestor. Unfortunately we have no access to detailed descriptions of mixed languages, so we cannot support this theoretical observation by an investigation of a real case.

The internal development of language *T* is caused by various changes which affected all parts of its system and, after a certain period of time, accumulated results of such changes make the new system so different from the original one that we have to talk about a new language L which is different from *T*. It is important to mention that a particular change never affects language development on its own, and only the accumulation of many changes causes a language disintegration.

The process of internal development of L can be represented as consisting of the following major stages:

stage I is L's crystallization: due to the accumulation of specific changes L's system becomes significantly different from the systems of its sister-languages;

stage II is L's internal evolution when L's system remains relatively homogeneous. As the changes of L's system continue to be accumulated, the differences between L and its sister-languages are constantly increasing;

stage III is L's disappearance caused either by its disintegration or by its 'death' (L or its daughter-languages are not used any more).

These three major stages of language development are not separated from each other by any sharp dividing line. To the contrary, in most cases, the transition from one stage to another is rather gradual and takes a reasonably long period of time. For example, if we talk about language disappearance due to its disintegration (the transition form stage II to stage III) we can expect that at first the differences between A, A' and A" (which would later develop into L's daughter-languages) were minimal (if any) and most changes were common to all of them, while later the changes would become language-specific. It is highly probable that speakers of A, for example, used forms of A' and did not identify them as foreign additions to their own language. Such borrowings often cannot be detected by comparative methods. That is why we have to assume that each stage is separated from another by a certain 'blind spot' whose features and duration are not quite known.

§9. The history of a language family is formed by the histories of individual languages. First its common proto-language goes through the three major stages of evolution (crystallization, integrated development and disappearance), then its daughter-languages go through them, then in turn the daughter-languages of these descendants, and so on. As a language always has only one ancestor and cannot

change it in time, this aspect of family evolution (its branching) has to be represented by a genetic tree of a fixed structure.

We distinguish between the notions of 'genetic tree' and 'evolutional tree'. A genetic tree is only a presentation of a family structure, while an evolutional tree is of a more complex nature to be discussed below.

As a genetic tree represents only linguistic continuity, for each language L we need to know only:

(i) L's direct ancestor (if any);

(ii) L's sister languages (if any);

(iii) L's daughter-languages (if any).

The following formal features of genetic trees reflect our understanding of a family's evolution:

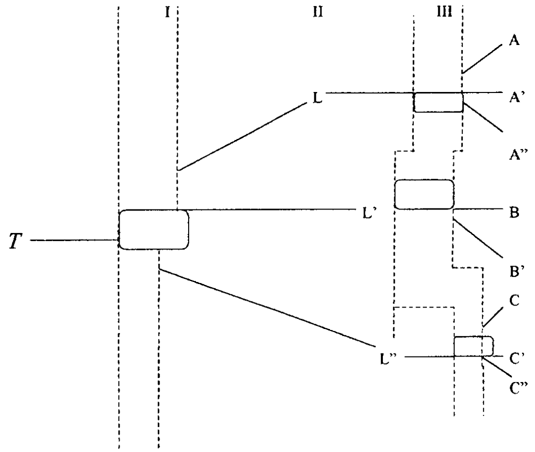
1. A genetic tree represents the internal structure of a family and thus includes only genetically related languages.
2. A genetic tree is formed only by nodes and directed arcs connecting these nodes.
3. There are two types of nodes:
   (i) those representing recorded languages;
   (ii) those representing entities postulated for the needs of classification.
   Often, but not always, these entities represent reconstructed languages postulated in the process of the family's investigation.
4. A directed arc represents the linguistic community connecting ancestor language *T* with its daughter-language L.
5. In each tree there is only one root node with no entering arc corresponding to the common proto-language of the family. This reflects the assumption that all related languages have developed from one common source - the proto-language of the family.
6. Apart from the proto language, every other language of the family always has only one direct genetic ancestor. In a genetic tree, every node, other than the root one, has only one entering arc. No node can have more than one entering arc.
7. A node without any arcs going out represents a language without known descendants; it could either be a language with no speakers (= a dead language) or a language without a significant dialectal diversity.

8. There is no limitation on the number of arcs going out of a node: none, one, two, or many. No well grounded reasons are known to support the idea that a genetic tree should always have a binary structure[8].
9. If two nodes are connected by an arc, this connection remains unchanged and under no circumstances do we accept a situation when, at one chronological stage, a node is connected to one ancestor, while in a later period, it is connected to another one. This reflects our fundamental belief that a language never changes its genetic affiliation.

It follows from the above that in order to create a genetic tree, we need to define how to identify nodes and how to connect them with arcs. As differences between nodes are caused by differences of accumulated changes rather than individual changes, there is no logical necessity to investigate these particular changes individually.[9]

§10. Each stage of language development has its own duration which is not necessarily identical across the family, and at any given period of time related languages can represent various stages of development (see Figure II).

Figure II



[8] Binary branchings in classifications often reflect not the internal structures of families, but rather the properties of classification procedures.

[9] This claim contradicts the generally adopted technique of classification with the help of innovations.

In this Figure Stage I covers the disintegration of ancestor language *T* and the formation of its daughter-languages (L, L', L"). In this period, the daughter-languages start to develop independently, and all common features adopted by them during and after this period are caused by borrowings, convergences, and other non-genetic reasons. Stage II covers the periods of integrated independent development of languages L, L' and L". Stage III covers the period of disintegration of the daughter-languages: L develops into A, A' and A", L" develops into B and B' and so on.

§11. Various changes in a language do not occur spontaneously, and the following explanation can be suggested.

Knowledge of a language is, first of all, an ability to express a given meaning in a text (verbal or written) and to extract all possible meanings from a given text. From this point of view, language is a device which matches meanings and texts (Mel'chuk 1988, p. 47). Speakers, however, also know the norms of the language usage adopted in their community. These norms allow them to make choices within options provided by the language: how to choose one of the synonyms, how to pronounce a particular sound, which morphological or syntactic structure is preferable in this situation, and so on. These norms are not necessary identical for all its members (cf. sociolinguistic variations within a community), but if a member speaks 'properly', he or she always knows the norms and follows them.

Different communities which use the same language do not normally share the sets of such norms and their members are often aware of that ('We do not use this word', 'This is the American pronunciation', and so on).

All changes in a language are always triggered by changes in the norms of usage. At first, a new norm only suggests the preferable choice among several options provided by the language, but after a certain period of time, this choice becomes the only acceptable possibility which leads towards changes of the corresponding linguistic structures.

Norms of language usage are in a constant process of change which in turn causes changes in the language structure, which accumulate such changes and results in significant differences between the various stages of the language development.

The acceptance of norms and their changes is crucial for language development. If various groups within the community began to adopt different norms this could lead towards the development of differences in the language they use: different norms trigger different linguistic changes and eventually would end up with the language's disintegration and formation of several new languages, each being a daughter-language of the same original one.

Therefore we can postulate the following connection:

change of norms of usage => changes of language's features => family evolution.

Of these, only family evolution is of a universal nature, while change of norms are community specific and changes of a language's features are language specific.

Therefore we have to assume that development of a language family is not just a simple sum of histories of individual languages. It can be seen as formed by three different processes:

1. Family evolution, which is a universal process of branching: an original language disintegrates into new ones, each being its daughter-languages.
2. Histories of individual languages formed by numerous changes affected systems of these languages. These changes are usually language-specific, but sometimes similar changes can be found in different languages. It is important to mention that a particular change in a language's history does not affect the family evolution. The statement, 'if a change *a > β did not occur we cannot talk about the split of languages A and B', is not correct.
3. History of norms of language usage within a particular community are always community-specific.

Here we will discuss only problems related to a family's evolution.

§12. When discussing a genetic classification, linguists usually distinguish two problems:

(i) identifying the genetic affiliation of languages: Germanic, Sino-Tibetan, Nostratic, etc.

(ii) creating an evolutional tree for a family.

When working out an evolutional tree for a family, we must first identify which languages are genetically related and form this family. To do that one can use the following formal procedure.

Two languages A and B are genetically related if the three conditions are fulfilled:

<u>Condition I: Existence of similar morphemes:</u>

(i) Genetically related languages A and B always share a sufficient number of similar morphemes[10].

If the two languages reveal a sufficient number of similar morphemes, one can assume that this similarity is not accidental, and it can probably be interpreted as evidence of genetic relation. Similar morphemes can be either lexical or grammatical, but the existence of similar lexical morphemes seems to be obligatory: there is no generally accepted language family for which languages do not share similar lexical morphemes, while for several well-established language families of Southeast Asia (Kadai, Vietic, Lolo-Burmese) no grammatical morphemes are known so far (Peiros 1998). If similar morphemes are not found, we do not have data for a further discussion of genetic relationship of these languages.

<u>Condition II: Genetic reasons for the similarities between morphemes:</u>

(ii) Sufficient number of similar morphemes in languages A and B belong to the core lexicon.

Similar morphemes can be found in all parts of the lexicon, but if the languages are genetically related, they always share morphemes from the core lexicon. We will develop and discuss this notion later (§19). Here it is enough to say that core lexicon includes words with simple universal meanings, which are less open to borrowings than other parts of a language lexicon. There is no doubt that words from the core lexicon can also be borrowed but the likelihood of borrowing here is usually lower. To the best of our knowledge, all known related languages always share words from the core lexicon. Thus one can conclude that if two languages share not only similar morphemes but morphemes which also belong to the core lexicon, it is more probable that these languages are genetically related.

Similarities between morphemes can be due to various reasons: common origin, borrowing, chance resemblance, and so on. To demonstrate the genetic

[10] For other views see, for example, Guy 1981, Nichols 1996.

nature of these similarities, we need a system of phonological correspondences between the languages:

Condition III: Existence of systemic phonological correspondences:

(iii) The phonological systems of A and B are connected by systemic phonological correspondences[11] with the element of one system corresponding to certain elements (one, several or none) in another.

(iv) The systemic phonological correspondences mentioned in (iii) are true for lexical similarities discussed in (i) and (ii).

These conditions are sufficient to provide us with formal criteria to judge if there is enough evidence to accept that two languages are genetically related and (due to transitivity of the notion[12]) that all languages related by them belong to the same linguistic family. It is important to mention that there is no additional requirement for grammatical similarities. However, where there are such similarities, they can provide an additional and often crucial support for a genetic claim.

The conditions discussed require that a set of systematic phonological correspondences be established between all genetically related languages.

However, to be able to establish such correspondences we need a certain level of understanding of the languages' relations. It would not be wise, for example, to try to obtain such correspondences investigating simultaneously English, Dutch, Russian and Polish. The more natural procedure would be at first to study separately Germanic and Slavic languages and then to compare the results. To do this we need, however, to be able to identify these two groups at least at the level of intuition. The whole process thus can be represented as consisting of four consecutive stages:

(1) a language family is identified using various heuristic procedures;

(2) a hypothesis about its classification is suggested;

[11] A phonological correspondence is a systemic one if it brings together reflexes of a particular proto phoneme. By the definition, a correspondence based on borrowings cannot be a systemic. A systematic correspondence can be either regular (e.g. found in many examples) or not.

[12] If language A is genetically related to language B and language B is genetically related to language C, then A is always genetically related to C.

(3) the formal method of comparative reconstruction is applied to the languages intuitively included in the family;

(4) a classification based on the results of the reconstruction is suggested.

The differences between stages 1/3 and 2/4 affect the reliability of the results and it is quite possible that some of the decisions made at the intuitive stages 1 and 2 will be rejected by the strict procedures of the analysis conducted at the stages 3 and 4. At the same time, one cannot start an investigation just from the third stage ignoring all intuitively based decisions.

Let us now limit our discussion to stage (4) which means that we are dealing with a well-established language family.

§13. A language family is formed by processes started in the past but with results observed in the present time. Studying these processes, linguists try:

- to reconstruct the common proto-language of the family and thus to explain the origins of structures of recorded language;
- to suggest a model of family evolution from its proto-language into historically attested languages.

The reconstruction of the common proto-language is conducted in several steps. It starts with the investigation of a group of recorded sister-languages and with a reconstruction of their proto-language. At the next stage of research, the same procedure is repeated, but instead of the recorded languages, we work with their reconstructed proto-language which is compared with its sister-languages, either recorded or reconstructed. Their ancestor is reconstructed which in turn, will be later used for more ancient reconstruction. The theoretical requirement is, that a daughter-language should never be used instead of its reconstructed ancestor: as soon as the reconstruction is completed, we should act as if more recent stages of its development do not exist at all (= what is not included in the reconstruction did not exist in the proto-language).[13]

The procedure of reconstruction thus goes in the opposite direction to the real process of languages development: starting with present recorded languages or

[13] We are free, however, to modify our reconstructions to include previously unaccounted features of daughter-languages.

relatively recent recordings, it moves back in history discovering at each step more and more ancient ancestors of these languages. This retrospective approach is the only justifiable approach in comparative linguistics.

A creation of an evolutional model follows the same pattern:

- we move in the direction opposite to the real process: from the present to the past;
- at any taxonomic stage S, we identify sister-languages A, A', A", which are all daughter-languages of L;
- at the chronologically preceding and thus more ancient taxonomic stage S_1, a search for languages specifically related to L (L', L", etc.) is conducted and their common proto-language *T* is identified. This is done with no reference to the situation at stage S (A, A' and A" or to daughter-languages of L' or L").

§14. At any period of its evolution, a language family is characterized by various degrees of differences accumulated in its languages. Therefore we need to differentiate two aspects of family evolution: a chronological aspect (languages exist in time and in any given period of time where they are at a particular stage of their development) and a 'divergent' aspect (the increase of differences between the languages). Ideally, any model of a family evolution should:

- be chronologically correlated, telling us about the relative chronology of languages of the same or different taxonomic levels ("Did the split of Proto Germanic take place much later than the split of Celtic?"). Chronological information, at least a relative one, seems to be an essential part of any model dealing with processes, including language development.

- provide us with information about diversity among the related languages which range from the cases where it is hard to decide if we are dealing with dialects of the same language or with two closely related languages to the cases where very complex research should be undertaken to demonstrate the very fact of the relationship. In the former case, the systems are quite similar while in the latter, only obscure traces of the similarities can be identified. Dealing with various groups of related languages, linguists always want to know how different the languages are within a group in comparison to another, usually better known group: "Are Eastern Slavonic languages closer to each other than the Southern Slavonic ones?", "Are Zhuang-Thai languages less genetically diverse (closer to each other) than Slavonic ones?".

These considerations suggest that an ideal model of a family evolution should represent:

- the structure of the family (= its genetic tree);
- the family's internal diversity;
- chronological stratification of the family's evolution.

The problems of chronological stratification belong to the theory of glottochronology and are not discussed here (see Starostin 1989). This leaves us with a partial model dealing only with branching and degrees of similarities.

§15. The development of a family structure, i.e. its branching, is connected to the notion of language continuity and thus the corresponding component of the model has to be based on pure genetic data.

Internal diversity of a family is determined by the accumulated amount of both internally and externally triggered changes. So evaluating these degrees we have two options:

(i) either to use the whole range of data, including similar loans, regional features, and other non-genetic features;

(ii) or to concentrate only on genetically caused features preserved by the languages from their common ancestors.

The second option (orientated towards genetically caused similarities) opens a possibility to use the same type of data as for modelling branching.

§16. The process of evolution of a language family based on the split of an original single language into its descendants is a universal one and can be observed in all language families regardless of the internal organisation of languages, the specific features of speech communities and other circumstances. Therefore it should be possible to suggest a universal model of evolution applicable to any language family of the world.

For example, we should be able to analyse in the same way the Vietic family of Austroasiatic (known for its lack of morphology and intensive contacts with the other languages of the Southeast Asian linguistic area), the Paman family (a typical Australian family both in its structural features and cultural and sociolinguistic characteristics of its speakers), and the Slavonic languages of Indo-European. Obtaining evolutional models of these three families should enable us to compare the relation between Vietnamese and Arem of Vietic to that of Yir-Yoront and Jabugay of Paman or Bulgarian and Russian of Slavonic. It should also enable us to answers questions such as: 'which of these pairs represents sister-languages?', 'which pairs are more diverse?'.

By studying a family's evolution, we model a real historical process. Our model therefore has to meet at least the following conditions:

I. As it deals with genetic relations between languages, it has to be based on pure genetic considerations and thus:

- criteria used in the classification have to be of pure genetic nature;
- these genetic criteria have to be made explicit.

II. The procedure used should be universal and applicable to any language family, regardless of their typology and other characteristics; so:

- the procedures must use features found in all human languages;
- the treatment of these features should be identical for any language family.

Therefore the evolutional trees are to be comparable across language families.

III. An evolutional tree generated by the model represents real historical relations between languages and thus:

- it should provide information about:
  (a) the structure of the family (= its genetic tree)
  (b) the internal diversity of the family;
- a procedure of classification has to be automatic and free of any type of personal preferences of scholars involved:
  (a) two scholars working independently should produce the same tree;
  (b) the choice of parameters for classification should be based on objective rather than subjective criteria;

  (c) the results of a classification should be open to a formal procedure of evaluation.

IV. The method of classification should be reasonably simple.

§17. Let us discuss now how to meet these requirements and to create an evolutional model.

The genetically caused similarities between languages A and A' or B and B' of Figure II are due to the retention of features developed in periods I and II, while their differences are the result of their independent development in period III. In the same way the genetically caused similarities between A and B or A' and B can only develop during period I, while their differences have appeared in periods II or III.

Theoretically one can expect to find that:

- the total amount of genetically caused similarities between A and A' should always be higher than between A and B or A' and B', as the existence of A and A' as a single entity was longer and thus more features were adopted;
- some of the genetically caused similarities between A and B or A and B' should be more or less identical: having retained from *T*, the common ancestor of the family and thus having a good chance of being retained with the descendants;
- the amount of genetically caused similarities between any pair of languages reflects the level of their relationship and thus their position in the genetic tree.

Here we would expect to hear the following argument. "Imagine, that language A, in the process of its internal development (period III), has lost all features inherited from L. In such a case the amount of its similarities with A' would not be different from that with B and thus it would be impossible to demonstrate its specific relation with A'. That is why the above mentioned considerations are not convincing". Such an argument is one of the numerous linguistic myths beseiging lexicostatistics. There are no proven records of such developments, i.e. information about well-defined language groups where languages have no traces of a specific common origin. The standard procedure of comparative linguistics would not be able to detect them and justify that A and A' should be kept together. Without common features, no substantial comparative claim about specific relations can be made[14].

Comparing the systems of *T* and its daughter-language L one finds:

(i) R(etentions) - features retained in L from *T* (result of drifts);

(ii) A(dditions) - features added to the system of L in the process of its crystallization or integrated development (result of additions);

(iii) S(ubtractions) - features of *T* not retained in L (result of losses).

Let Σ(L) be the set of structural features of L. By definition Σ(L) includes different features of L: its phonemes, morphemes, grammatical and syntactic rules and so on. As we can separate L from all other languages, including its direct ancestor *T*, Σ(L) is always different from Σ any other language.

Let R(L,*T*) be the set of features retained in L from its ancestor *T*. This R(L,*T*) is always smaller than Σ(L), as it also includes A(L) - a set of specific additions to L not found in its ancestor. Therefore Σ(L) = R(L,*T*) +A(L).

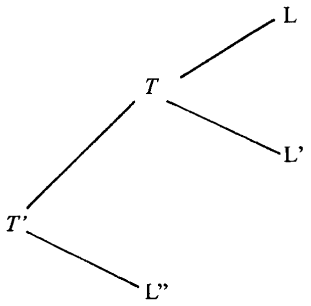
If L and L' are sister-languages, they retained features of their common ancestor *T* and those features are the only source of genetically caused similarities found in the two languages. L and L' have developed separately and thus their losses of the original system cannot be identical. Some of them could coincide, but the total set of losses would never be the same. As two languages L and L' can never develop in exactly the same way, each has language-specific changes and R(L,*T*) is always different from R(L',*T*).

Let D(A,B) - we will call it '**genetic distance** between A and B' - be a set of genetically identical features retained in A and B. For L and its direct ancestor *T* D(L,*T*) is identical to R(L,*T*). For two sister-languages L and L', their D(L,L') is an intersection of two sets: R(L,*T*) $\cap$ R(L',*T*).

For any L, its D(L,L') is always smaller than its D(L,*T*), but is always greater than D(L,B), where B is not L's direct ancestor or a sister-language. This claim is based on the following arguments. For the family represented in Figure III we will always find more similarities between L and L' than between L and L" or L' and L". Shared features of L and L' are inherited from *T*, while shared features, say, between L and L" are inherited from *T'*. L" and *T* are sister-languages and their sets of shared features are always smaller than the sets of features inherited by each of them from their common ancestor *T'*. L and L' are daughter-languages of *T*, which is the source of all their common features. This source, as we have already discussed, is formed by retentions from the previous stage as well as by specific additions not known in its sister languages: R + A. This unique set is the original source for L and L'. It is absolutely impossible that one of these languages would lose all the features specific to its ancestor adopting at the same time some features from the ancestor sister-language L". Therefore D(L,L') is always smaller than R(L,*T*) or R(L,*T'*). As D(L,L") is in turn always smaller than R(L,*T'*) it should be also smaller than D(L,L').

[14] A discussion of 'retention rates' is given below (§ 26).

Figure III
(same as Figure I)



If we are now able to suggest a procedure to identify and compare these genetic distances we should be able to answer the following questions:

(i) what languages are directly related?

(ii) which of them are sister-languages?

(iii) what language is the common ancestor of the family?

This procedure will provide us with a formal tool to create genetic trees and to represent the internal diversity of the family, thus achieving (apart from getting absolute datings) objectives of the evolutional model.

§18. To evaluate genetic distances we can either use the all the information available for these languages or choose certain sets of parameters which would allow us to achieve the same goal in a more economic way. The second approach seems to be preferable, but raises two questions: 'how to select the parameters?' and 'how to use them to obtain a classification?'.

Let us first discuss how to select the parameters. The evolution of a language family is caused by various changes whose accumulated results led to the disintegration of a former single language. Therefore it seems logical to choose parameters related to such changes.

We have discussed already the difference between internal and external changes. Of these only internally caused changes can be used to investigate genetic relations between languages. External changes are caused by foreign influence and thus they cannot indicate pure genetic relations. Rather they can be used in classifications whose goal is to identify linguistic areas ('Balkan languages', 'Southeast Asian languages') or groups of languages under a common influence ('languages of Sino-centric world'), but not a genetic relation. Therefore for each change (or its result) used in any genetic model we need to provide evidence of its internal nature. Such evidence presented explicitly must always be included in the data.

There are two main ways to select diagnostic changes:

- to create a fixed list of them to be applied to various languages; or

- to suggest a formal procedure of their selection, which would provide us with different lists depending on the family to be investigated.

As it is highly unlikely that a change can be found in all human languages, it seem preferable to deal with a selection procedure, rather than with a certain list, knowing that such a list is by no means universal.

The selection of diagnostic parameters determines the whole model, and should reflect the requirements outlined above:

- selected parameters are to be found in all human languages;

- the parameters are to be either internally caused changes or the reflexes of such changes;

- selection of parameters should always be conducted automatically, i.e.; not affected by the personal preferences of a researcher: 'I think that this change is revealing, but I cannot prove this'.

This leads us to the question of 'what kind of parameters can be chosen for the model?' We have decided already that the parameters chosen have to be either internally caused changes or their reflexes. However it seems also important to use only free and not triggered changes. The reason for such a limitation follows.

Imagine that we are dealing with a language *T* which distinguishes ten noun cases. The morphs for three of them - Accusative, Genitive and Dative - differ only by final stops: *at, ap* and *ak*. In the history of its daughter-language L final stops were lost which triggered the loss of distinctions between these three cases and required restructuring of the whole case system. In L', another daughter-language of *T*, final stops are preserved and the original case system remains unchanged. If we now choose the case markers to be our parameters, we would most probably come to a wrong conclusion. In this example, only the phonological change can be selected as a parameter to be used in the model, provided that it has not been triggered by other changes.

§19. In any evolutional model, we are dealing with genetically related languages, which means that all of them are daughter-languages of the family proto-language *T*. This means that there is always an uninterrupted language continuity between *T* and its daughter-languages. We have discussed already (§4) language continuity defined with the help of core meanings (over the whole period of language continuity, the core information is conveyed mainly with the help of linguistic means of the same origin).

> "As is well known, natural language has only two major types of expressive means—lexical and non-lexical—to encode the information a sentence carries. Lexical means are simply words. The set of lexical means used in a sentence is the list of all wordforms, or lexeme occurrences, that constitute it.
> Non-lexical means are of three varieties:
> • linear order of wordforms
> • prosody (intonation contours, pauses, phrase and sentence stress)
> • inflections (i.e. morphological categories).
> Of these three varieties, word order is the most important and the most universal, being necessarily present in every language and in every sentence. It is imposed by the physiologically conditioned linearity of human speech. Inflections, on the other hand, are the least important ... because they are the least universal linguistic means: some languages lack them entirely; in all

languages there are uninflected words, which, however, are syntactically linked to other words in a sentence. Prosody occupies an intermediate position.
There are no other types of linguistic expressive means.
Both lexical and non-lexical expressive means of a language can be used in one of the following two ways:
- Either in a SEMANTIC capacity, i.e., to convey meaning immediately...
- Or in a SYNTACTIC capacity, i.e. to mark relations between linguistic entities..." (Mel'chuk 1988, p.19-20).

Usage of lexical means in a syntactic capacity or non-lexical means in a semantic capacity is always language-specific. Therefore, if we are planning to use expressive means of languages to build a universally applicable model of a family's evolution, we can employ only lexical means in a semantic capacity and/or linear order of word forms used in a syntactic capacity.

Any model of a family's evolution establishes continuity between languages which has been defined with the help of core meanings, so is seems natural to build a model also orientated towards semantics, or more precisely to the lexical means of languages used in their semantic capacity. Otherwise we will adopt a logically inconsistent approach with language continuity defined with the help of core meanings, while other features are used to identify this continuity[15].

Core meanings are found in any human language where they are normally represented by various lexical means of this language, words or more complex structures. These lexical means form the **core lexicon** of the language. As there is no apparent reason for these core meanings to be dropped from the language, we can assume that the corresponding words have a better chance of being preserved than words from other parts of the lexicon. At the same time, it is well known that words from the core lexicon can also be borrowed or substituted due to internal development. In all such cases, the corresponding meanings nevertheless remain in the language.

The core lexicon of language A is a set of words or more complex lexical structures whose meanings are the most precise equivalents of the core meanings. The core lexicon of language B is another set of words which represent the same set of meanings. These two core lexicons are not identical but we can expect that for every word from A, there is one or several words in B whose meanings are

[15] It is extremely important to mention here, that we do not claim that other features (like genetically caused irregularities of verbs in Germanic languages) do not indicate specific genetic relations. The claim is that a universal model should be built on an analysis of core meanings because of their universal nature.

identical and represent a particular core meaning [. These words are [-synonyms. As the core meanings are supposed to be universal, it is theoretically possible to identify the core lexicon in any human language. All such core lexicons would include forms with comparable meanings.

§20. We have already formulated the three basic logical assumptions of the evolutional model:

(1) any two genetically related languages have common features which are retained in these languages from their common ancestor[16];

(2) these common features can be found both in the lexicon and grammar (in the wider sense) of these languages. There is, however, no well established language family without common lexical similarities;

(3) there is always a correlation between the position of languages in a classification and the amount of genetically caused similarities between them. If two languages A and B are grouped together, they always retain more common features than with any other, more distantly related language, C. If A, B and C are equally distant in a classification, no pair of them reveals a significantly higher amount of common inherited features;

To these, we can add now a fourth assumption:

(4) the same correlation is observed if we compare the core lexicons of languages: if languages A and B are closer to each other than C, more words of common origin are found in their core lexicons, and fewer are shared by A and C or B and C.

Now we can formulate two basic postulates of lexicostatistics:

1. Core lexicons can be used to identify genetic distances between related languages and to model the evolution of the whole family.

2. It is possible to do this with the help of a limited sample chosen from the core lexicons of related languages.

§21. We have already discussed how a language's core lexicon is formed by words of this language which represent the core meanings. These meanings are,

[16] An absence of such similarities indicates the absence of transparent relationship which, however, does not rule out that the languages may be distantly related.

by definition, universal, while the corresponding words are language-specific. It follows from the proposed assumptions that it should be possible to compile a list of core meanings suitable for an evolutional model. Such a list, let us call it **‘σ-list’**, should have the following features:

1. it is formed by core meanings which, by definition, are usually represented in all human languages regardless of habitat, economic activities or other characteristics of their speakers;

2. it includes easily identifiable meanings which are normally associated with simple words in languages. From this point of view, σ-list should not include such words as ‘life’ or ‘anger’, as for many languages it would be quite difficult to identify corresponding words;

3. the chosen meanings should be free from taboos and other cultural restrictions on their possible usage. Thus we cannot include in the list such core meanings as ‘to give birth’ or ‘excrements’;

4. meanings included in σ-list should be historically independent from each other and a change in one meaning should not trigger changes in other meanings from the same list. From this point of view, the list is not supposed to include meanings from the same semantic field, like ‘eye’, ‘pupil of the eye’, ‘white of the eye’, and others, as a change in one of them can easily cause a domino-like effect for changes among the others.

If these conditions are met, it is not very important how many meanings are included in the σ-list. However, a very short σ-list (say 20 items) is not a good option, as the impact of each entry becomes too great. On the other hand, an analysis of a very long list (say 1,000 items) can be excessively time consuming and thus is not convenient. It is important to mention, however, that differences in σ-list can significantly affect the evolutional tree obtained, so that one cannot compare results obtained with the help of a 200-item list with the results obtained with the help of 100-item or 35-item lists. Even differences of lists in 10 - 15 meanings may make the trees incompatible.

Figure IV
Standard α-list

1. all (as in ‘all of the stones’)
2. ashes (cold ashes left after fire)
3. bark (of a tree)
4. belly (the outside part of human belly)
5. big (big in size, as opposed to small)
6. bird
7. bite (bite as in eating)
8. black (black colour)
9. blood
10. bone
11. breast (female breast)
12. burn (vt., to burn sticks)
13. claw or fingernail
14. cloud
15. cold (as water)
16. come
17. die (of a person)
18. dog
19. drink (as water)
20. dry (as clothes)
21. ear
22. earth (as in ‘a shovel of earth’)
23. eat
24. egg
25. eye
26. fat (n., as in ‘meat and fat’)
27. feather
28. fire
29. fish (n.)
30. fly (as a bird)
31. foot
32. full (as a basket)
33. give
34. good
35. green (colour)
36. hair (of head)
37. hand
38. head
39. hear (as sounds)
40. heart (as body part)
41. horn (cow’s horn)
42. I
43. kill (as a person)
44. knee
45. know (as in ‘I know his’)
46. leaf
47. lie (as ‘the boy lies on the floor’)
48. liver
49. long (rope)
50. louse (hair)
51. man (as opposite to woman)
52. many (as in ‘many stones’)
53. meat (as in ‘to cook meat’)
54. moon
55. mountain
56. mouth (of a person)
57. name (of a person)
58. neck (as opposite to throat)
59. new (as in ‘new clothes’)
60. night (as opposite to day)
61. nose
62. not (as in ‘not new’)
63. one
64. person (= human being)
65. rain
66. red (colour)
67. road or path
68. root (of a tree)
69. round (object)
70. sand
71. see (as in ‘I can see him’)
72. seed (of a fruit)
73. sit (as in ‘he is sitting here’)
74. skin (of human being)
75. sleep (as in ‘he is sleeping’)
76. small (in size, opposite to big)
77. smoke (of a fire)
78. say or speak (an in ‘he is saying something’)
79. stand (as in ‘he is standing there’)
80. star
81. stone
82. sun
83. swim (as in ‘he is swimming there’)
84. tail (of a dog)
85. that (far from here)
86. this (close to here)
87. thou (2nd Sg.)
88. tongue (of a human being)
89. tooth
90. tree
91. two
92. walk or go (as in ‘we will walk there/ here’)
93. warm (as water)
94. water (fresh water)
95. we (Pl. excl.)
96. what?
97. white (white colour)
98. who?
99. woman (as opposite to man)
100. yellow (yellow colour)

There are many ways to compile a σ-list, but we prefer to use the standard one which includes 100 meanings suggested by Swadesh, despite the fact that the list is not ideal as:

- some of its meanings are not universal: for example, until recently the meaning ‘horn’ was not known in Australian Aboriginal languages;

- not all of its meanings are completely independent: as, for example, in the case of 'bark' and 'skin';
- not always are the meanings free from cultural impact: it is possible, for example, the meaning 'person' which often has connotations like 'us, true human beings'.

Nevertheless it seems convenient to deal with the standard list, as all other σ-lists always have problems of their own. One advantage of the standard list is that it has already been used for many languages, which helps in collecting data and comparing results.

§22. The actual lexicostatistical procedure starts with the completion of the **diagnostic lists** (σ-lists) of all the languages under investigation.

A σ-list of language L is formed by words or more complex lexical units, whose meanings are the most precise unmarked equivalents of corresponding σ-meanings. It represents this language as it is used by a certain well-defined group of people in a particular period of time (it is dialectally and chronologically specific). Therefore it always has to include only forms taken from a certain dialect rather than from several different dialects or even closely related languages. It also has to represent a particular chronological period and not include forms used in different periods of the language's history.

A word included in a σ-list has to be the most precise translation of the σ-meanings into the language and thus be:

(i) the most common representation of a σ-meaning in the language. We have to include only the most widely used word, as for example, the first word given for an English-L dictionary entry. Such dictionaries are often the best source of data for compiling a σ-list if a judgment of native speakers is not available.

(ii) taken from the unmarked, neutral style or register of the language, as otherwise the word would not be the most precise representation of a σ-meaning. If for example, word traj means 'tail', but is used only by hunters, while other speakers use another word kan, only the latter is to be included in the σ-list.

Taboos raise a specific theoretical problem. In taboo situations a particular word can temporarily not be used and another word used instead. Over a certain period of time, the original word would regain its unlimited role. However, to observe a taboo, a speaker is supposed to know both words (tabooed and its substitution), but to use only the substitution. We may include both forms in a

σ-list, but it seems more logical to deal only with the original word, treating its taboo substitution as a less general, marked form with no reason to be included in a [σ-list. It is worthwhile to mention, however, that clear taboo substitutions have not been encountered in our work with various language families, including the Australian Aboriginal languages, known for their taboo replacements (Dixon 1980:28; Alpher & Nash, forthcoming). All languages' dictionaries usually provide us mainly with main words, while taboo substitutions, if included, are always specifically marked.

(iii) The process of the completion of a σ-list should not be affected by any historical considerations, and every σ-list should represent the typical usage of the period chosen. It means that we should not include in a σ-list forms which have good etymologies, but are not central from the point of their usage. For example, despite the fact that archaic Russian *oko* 'eye' has a good Indo-European etymology, the Russian σ-list includes only the modern form *glaz* .

(iv) Sometimes a language does not have one single word whose meaning corresponds precisely to an σ-meaning. Two different situations are represented here:

A language does not have a word with a σ-meaning, instead a word with a broader meaning is used. Russian, for example does not distinguish between the meanings 'leg' and 'foot'. As the corresponding word - *noga* - is the most precise representation of the σ-meaning FOOT, this word is included in the Russian σ-list.

Nyawaygi, an Australian Aboriginal language, does not have a single word which represents the σ-meaning MOON; instead it has two different words palanu 'new moon' and ıilkan 'full moon' (Dixon 1983). Both words should be included in the σ-list of Nyawaygi, thus achieving the aim to have the corresponding σ-meaning be represented completely.

If these requirements for the selection of candidates to a σ-list are met, these properly chosen words are called (proper) **list-members.**

σ-lists of various languages have different sets of forms, but they always include forms which represent only one hundred core meanings. This fact is used when we compile **lexicostatistical tables.** A lexicostatistical table includes all σ-synonyms (one or several words) found in all the languages analysed. Table HAND, for example, includes forms of languages which are synonymic and mean 'hand'. This meaning can be represented by a word which means only 'hand', by a word which means 'hand / arm', by a simple word, a compound or even by several words. Each lexicostatistical table has two columns: one lists the languages under investigation, while another lists the forms of these. Each line of a table thus connects a language with its form. As [-list is formed by one hundred meanings, the whole database of the analysis includes one hundred lexicostatistical tables.

The lexicostatistical table 'STONE' for several Central Pacific languages[17] of the Austronesian family is:

Languages	Forms
East Fijian	βatu
West Fijian	βaču
Rotuman	hɔfu
Mele-Fila	fatu
Tahitian	ʔoofaʔi
Rapanui	maˈea
Nukuoro	hadu
Maori	koohatu, poohatu
Samoan	maʔa
Tongan	maka
Hawaiian	poohaku

§23. The next step of the procedure is the etymological identification of forms presented in individual lexicostatistical tables. For each word J of language L included in table T, we have to give Yes /No answers to two questions:

- is β a borrowing?

- does β have the same origin as form γ of L' in the same table?

Two comments are needed here:

(i) the etymological identification of the forms given in table T is not a complete etymological analysis. The aim of the latter is to find all cognates preserved in the languages, including those which changed their meanings: 'meat' => 'bird' or 'eye' => 'face'. Such forms with different meanings can sometimes be included in several different lexicostatistical tables or be left outside the investigation. Etymological identification in lexicostatistics deals only with data included in one particular table T and discusses only one question: if forms β and γ

[17] Lexical data is taken from Tryon 1995 and several language dictionaries.

of languages A and B included in T are cognates, or in other words, whether both J and K represent independent uninterrupted developments of the same *T* of their common ancestor. From this point of view, the fact that the word 'water' of language A is a cognate to the word 'rain' of language B and to the word 'cloud' of language C is irrelevant, as these words are included in three different lexicostatistical tables.

(ii) in the procedure, we analyze only lexical morphemes of the words included in table T. Affixes are not analyzed. This restriction follows from the orientation of the whole model towards lexical rather than grammatical means of a language.

Etymological identification is supposed to be based on detailed knowledge of the historical phonology and etymology of the family we investigate. Without it, one cannot conduct reliable identification of borrowings and cognates and thus full lexicostatistical analysis is not applicable[18].

After etymological identification is conducted, table T is given a the third column: information about the origins of all lexical morphemes included in it. This information is represented in numerical form:

- negative numbers for loans

- positive numbers for original words.

Words of the same origin have identical (positive or negative) numbers.

The etymologized lexicostatistical table 'STONE' for languages given above is[19]:

Languages	Forms	Etymological information
East Fijian	βatu	1
West Fijian	βaču	1
Rotuman	hɔfu	2
Mele-Fila	fatu	1
Tahitian	ʔoofaʔi	3
Rapanui	maʔea	4

[18] One can, however, use 'Preliminary lexicostatistics' as a heuristic method (see Peiros, in this volume).

[19] Etymological information is taken from Biggs MS.

Nukuoro	hadu	1
Maori	koohatu, poohatu	1
Samoan	maʔa	4
Tongan	maka	4
Hawaiian	poohaku	1

The forms identified with the number 1 are reflexes of Proto Austronesian **batu*; the forms identified with the number 4 are reflexes of Proto Polynesian **maka*; the forms identified with the numbers 2 or 3 are isolated in the table.

The same table for several Mon-Khmer languages is:

Languages	Forms	Etymological information
Katu	dəl	1
Bruu	təmaw.L	2
Kui	təmau.L	2
Pakoh	bul	3
Wa	si.mauʔ.B	2
Lawa	səmoʔ	2
De'ang	mau	2
Plang	kaʔ.4 muʔ.2	2
U	mo.2	2
Khmu	klà:ŋ	-4
Ksinmul	ʔəliəŋ	-4

The forms identified with the number 2 are reflexes of Proto Mon-Khmer **Cəmauʔ* > Proto Katu **[t/d]əmhaw* (Peiros 1996 N 475) and Proto Palung-Wa **səmauʔ* (Peiros MS). The forms identified with -4 are loans from a Sino-Tibetan source: Proto Sino-Tibetan **Lə:ŋ* / **Lə:k* (Peiros & Starostin 1995, 3:69)

Further analysis is based on these etymologically supported numbers and not on the actual forms given in the tables. We do not need to know any more that forms in Table STONE are *stone* and *Stein* as in English and German or *maʔa* and

maka as in Samoan and Tongan. All we need to know is that in both cases we are dealing with genetically identical forms.

§24. The **lexicostatistical data-base** for a family is formed by one hundred lexicostatistical tables which include:

(i) forms from all the languages under investigation;

(ii) the etymological identification of each form included in a table.

The next step of the procedure is a statistical analysis of this data-base and the completion of a lexicostatistical matrix of a family. The basic idea of this procedure is the following: for each pair of languages, we calculate the percent of etymologically identical words and include them in a matrix.

Borrowings and situations with more than one form represented in a language require special discussion.

When dealing with borrowings, we can choose one of the following strategies:

(i) we can treat borrowings in the same way as original words. With such an approach, borrowings from the same source are to be analyzed as genetically caused similarities. This would affect the amount of similarities between the languages which in turn would affect their position in the evolutional model. As we have already decided that this model represents only genetic relations of languages and not their contact, cultural, typological or other relations, such an approach cannot be accepted;

(ii) we can agree to treat borrowings (regardless of the fact that they can be of identical origin) as a lack of genetic identity. This would lead us to the same problem as in (i), i.e. an effect of borrowings on a genetic classification;

(iii) borrowing is not a genetic process and its results - various loans, simply substitute the original words which are not any more available for our investigation. Therefore, it seems logical to treat loans as lack of information, rather than to include them in an analysis of genetic development.

Forms of two languages in a table can be related in three different ways:

(a) L		L'	(b) L		L'	(c) L		L'
α	⇔	γ	α	⇔	γ	α	⇔	γ
			β	‖	δ	β	⇔	δ

If we treat (c) as representing two separate cases of genetic identity: α ⇔ γ and β ⇔ δ, then for (b) we have to talk about one case of identity (α ⇔ γ) and one case of lack of identity (β || δ). It means that statistically (c) would be treated as 2 identities, (b) - as 0 (one identity + one lack of identity = 0) and (a) as 1 (one identity). Such an approach does not seem to be the best one, and for every pair of languages, we count only identity regardless of the fact that in reality we have more than one pair of etymologically identical morphemes. In the situation:

L		L’		L”
α	⇔	γ		
β	⇔	δ		
		ε	⇔	ζ

we will accept only one countable identity for L and L’ and one for L’ and L”.

§25. There are three types of relations between related languages:

1. Specific:
   1.1 the relation between an ancestor and its daughter-languages;
   1.2 the relation between sister-languages (the languages with the same direct ancestor)
2. Non-specific: the relation between other types of related languages.

It follows from the above discussion, that in a lexicostatistical matrix we expect to find that specifically related languages are marked by a higher percent of common shared words across the whole lexicostatistical data-base.

Let us examine the matrix in Figure V.

Figure V: A sample lexicostatistical matrix

	A	B	C	D	E	F	J
A		75%	40%	42%	40%	42%	38%
B	75%	-	40%	42%	40%	42%	38%
C	40%	40%	-	60%	50%	52%	50%
D	42%	42%	60%	-	52%	48%	50%
E	40%	40%	50%	52%	-	70%	68%
F	42%	42%	50%	48%	70%	-	72%
J	38%	38%	50%	50%	68%	72%	-

We can identify three groups of specifically related languages in this matrix: A/B, C/D and E/F/J. The percentages shared by languages of each of these groups are higher than for the languages across the groups:

	A	B	C	D	E	F	J
A		75%	40%	42%	40%	42%	38%
B	75%	-	40%	42%	40%	42%	38%
C	40%	40%	-	60%	50%	52%	50%
D	42%	42%	60%	-	52%	48%	50%
E	40%	40%	50%	52%	-	70%	68%
F	42%	42%	50%	48%	70%	-	72%
J	38%	38%	50%	50%	68%	72%	-

If so, the languages of every group have to have a common ancestor: L for A and B, L' for C and D, and L" for E, F and J. These ancestor-languages existed earlier than the period represented by matrix I. We can represent the relations between the three ancestor-languages in matrix II:

	L	L’	L”
L (A/B)	–	41%	40%
L’ (C/D)	41%	-	50%
L” (E/F/J)	40%	50%	-

Higher percentages between L’ and L” suggest that these two languages are specifically related and at an earlier stage of family’s evolution they were represented by a single common ancestor.

Comparing core lexicons of language L and its ancestor *T*, we can find the same picture as we have already discussed:

(i) words retained in L from *T*;

(ii) words known in L, but not in *T*. The words have appeared in L after it separated from *T* due to various additions to L’s lexicon.

The balance between groups (i) and (ii) reflects the level of similarities between these languages: more closely related languages always have more common words. This theoretical suggestion is based on our experience in comparative linguistics, as this balance can be observed for any pair of well studied languages. Therefore one can evaluate the genetic distance between language *T* and its descendant L on the basis of the amount of words retained in the core lexicon of L from the core lexicon of *T*.

§26. This suggestion contradicts the claims made in several substantial publications of Blust. According to him, the Austronesian languages have 'evident variability in retention percent' (Blust 1993:245), and percents of words retained by sister-languages from their common ancestor vary within 20-30 percent. If correct, this completely undermines the lexicostatistical method.

Let us, however, discuss Blust's arguments.

For many years, this scholar developed a classification of the Austronesian languages. This widely accepted classification has grouped the languages in the following tree:

Austronesian Family

1. Atayalic
2. Rukai-Tsoic
3. Paiwanic
4. Malayo-Polynesian (MP):
   i. Western Malayo-Polynesian
   ii. Central-Eastern Malayo-Polynesian (CEMP):
      a. Central Malayo-Polynesian (CMP)
      b. Eastern Malayo-Polynesian (EMP):
         South Halmahera - West New Guinea (SHWNG)
         Oceanic (Oc) (Blust 1978; Tryon 1995).

Blust has reconstructed the σ-lists for main proto languages of Malayo-Polynesian: PMP, PCEMP, PEMP and POc. Comparing these reconstructed lists with the σ-lists of various languages, he came to the following conclusion: "Almost all CMP languages, apart from those of the Aru Islands, are lexically quite conservative, with a mean retention percent of reconstructed PMP basic vocabulary of 38.9. In other words, the typical CMP language has a high concentration of lexical items that belong to cognate sets that are widely distributed in the Austronesian family. The SHWNG languages, on the other hand, are only moderately conservative (mean retention percent 25.6). The Oceanic languages vary widely in retention percent, from lexically rather conservative languages such as Ruga (39.5), Fijian (39.5), Trukese (37.8), Motu (36.7), Sa'a (36.2) and the Polynesian languages (ranging from about 33 to 40 percent) to lexically very innovative languages such as Jawe (19.1), Roviana (16.5), Misima (15.7), Kilivila (14.6), Teanu (10.8), Dehu (9.8), or Kaulong (5.7)" (Blust 1993,245).

Evaluating these conclusions, we need to address three main issues:

- the genetic classification of the Austronesian languages;
- the notion of 'retention rates'
- the procedure of evaluating these rates.

We cannot discuss here the whole problem of genetic classification of Austronesian (for an overview see Ross 1995), especially the question as to what extent the proposed Blust's classifications is a genetic one. To prove the genetic nature of a classification, one needs to provide conclusive evidence that all features used to justify the suggested branching are of pure genetic origin and do not represent regional convergence, borrowing or other non-genetic developments. In his classifications Blust follows the common technique of using selected innovations (phonological, grammatical, lexical) to prove the groupings. However, when dealing with an innovation, we always have a good chance that it be an externally caused and/or triggered change. Often, even for such well-known families as Slavonic, it is very difficult to prove the opposite and to demonstrate that a feature treated as an innovation is a free internally caused change. Austronesian comparative studies belong to the most developed areas of comparative linguistics, but still it is too early to believe that we can rule out the possibility of non-genetic origins of group-specific innovations, especially when such features are mainly losses or merges (see, for example Blust's list of MP innovations - Blust 1990).

The idea of retention rate, as we understand it[20], can be presented as the following: one can compare σ-lists of languages L and its direct ancestor and count the percent of words which are genetically identical in these two lists. This percent represents the retention rate of L.

[20] The original unpublished work of Blust dealing with the retention rates is not available to us.

Retention rates were establish for several pairs of languages (see, for example, Bergsland and Vogt 1962; Starostin 1989). All these test cases, however, are based on comparisons between recorded and well-known languages, which allows scholars to compile the σ-list with full confidence, choosing appropriate list-members.

It is much more difficult to compile a σ-list for a reconstructed language, than for a well known recorded one, as normally we do not have sufficient tools to prove that a word meets all the conditions required for a list-member: to be the main unmarked representation of an α-meaning, to belong to a particular dialect and at a chronological level and so on. In such cases, a word with a wider distribution can be seen as the proper list-member, which is not always true (see, for example, discussion of 'aging' of words in Starostin 1989) and additional research is needed to support that the most widely spread modern reflexes indicate the proper list-member of the proto language's σ-list. With many hundreds of Austronesian languages yet to be synchronically described, it seems to be practically impossible to substantiate sufficiently such claims[21].

Therefore we have solid reasons to believe that Blust's lists for various proto languages are not lexicostatistical σ-lists, but are lists of reconstructed forms whose reflexes in modern languages have wider distribution and are also represented in corresponding ancestor languages. This is supported by the following observation made by Blust about PMP, PCEMP, PCMP and POc: "The lexicostatistical comparison of the four protolanguages is of some limited interest. Systematic attempts to reconstruct Swadesh's 200-item test lists at various time-depths show clearly that Proto-Central Malayo-Polynesian was hardly distinct from Proto-Malayo-Polynesian (98 percent similar). A comparable relationship holds for Proto-Central Malayo-Polynesian in relation to Proto-Central-Eastern Malayo-Polynesian (96 percent). The comparison of other pairs of these protolanguages yields only moderately lower values: PCMP and PMP (94 percent), POc and PCEMP (93 percent), POc and PMP (88 percent), POc and PCMP (84 percent)". (Blust 1993, 245).

[21] In fact, due to the lack of developed reconstructions in Austronesian studies linguists are often dealing with 'comparisons' rather than with 'etymologies' (see Peiros, in this volume).

The percents given by Blust form the following matrix:

	PMP	PCEMP	PCMP	POc
PMP	x	98	94	88
PCEMP	96	x	96	93
PCMP	94	96	x	84
POc	88	93	84	x

Note that this matrix is based on 200-item list. For the standard list Blust's data allows us to build the following matrix:

	PMP	PCEMP	PCMP	POc
PMP	x	98	96	87
PCEMP	98	x	100	90
PCMP	96	100	x	
POc	87	90	89	x

It follows from this matrix that the reconstructed PCEMP list is identical to that of PC and nearly identical to the list for PMP. Accepting these results we have to conclude that no lexical changes have occurred over the whole period elapsed from PMP to PCEMP, which is quite improbable. It is better to assume that the matrix shows simply that forms reconstructed for PMP were preserved in Proto CEMP leaving completion of proper σ-lists for the time being.

If our understanding of the percents studied by Blust is correct, we can say that they, in fact, represent the extent to which languages confirm the suggested reconstructions, and they do not provide us with conclusive evidence of significant differences of retention rates among the Oceanic languages, and thus they do not contradict the theory of lexicostatistics.

§27. A lexicostatistical matrix includes sufficient data to model the evolution of a family. Percents included in it indicate directly degrees of diversity between languages, while their interpretation (not discussed here) provides us with the genetic tree of the family.

Here we cannot discuss the formal procedure of matrix interpretation (see Peiros and Starostin, in progress) which is based on the assumption that an evolutional tree can be obtained only from the whole matrix, rather than though analyses of percents revealed by individual pairs of languages. This procedure is used in STARLING, a software package designed by Starostin.

LITERATURE

Alpher, B. & Nash, D. (forthcoming) *Lexical replacement and cognate equilibrium in Australia.* 45 pp.

Bergsland and Vogt, 1962, *On the validity of glottochronology.*
In Current Anthropology 3: 111-153.

Biggs, B., 1991, *A Linguist revisits the New Zealand Bush.*
In Pawley, A., ed., Man And a half, Auckland: The Polynesian Society, pp. 67-72.

Biggs, B., MS POLLEX (computer print-out, University of Auckland).

Blust, R., 1978, *Eastern Malayo-Polynesian: a subgrouping argument.*
In S. Wurm and L. Carrington, eds., Second International Conference on Australian Linguistics: proceedings. PL, C-39: 181-234.

Blust, R., 1981, *Variation in retention rate among Austronesian languages.*
Paper presented to the Third International Conference on Austronesian Linguistics, Bali, Indonesia (not seen).

Blust, R., 1990, *Patterns of sound change in the Austronesian languages.*
In Baldi, P. ed. Linguistic change and reconstruction methodology.
Berlin, New York: Mouton de Gruyter, 231-263.

Blust, R., 1993, *Central and Central-Eastern Malayo-Polynesian.*
In Oceanic Linguistics, 32/2: 241-293.

Dixon, R., 1980, *The languages of Australia.*
Cambridge: Cambridge University Press.

Dixon, R., 1983, *Nyawaygi.*
In Dixon R. and B. Black, eds., The Handbook of Australian Languages. Vol. 3.
Canberra: Australian National University Press, 430 -531.

Guy, J.B.M., 1981, *Glottochronology without cognate recognition.*
In Pacific Linguistics, B-79.

Mel'chuk I., 1988, *Dependency Syntax: theory and practice.*
N.Y.: State Univ. of New York Press.

Nichols, Johanna, 1996, *The comparative method as heuristic.*
In M. Durie and M. Ross. The comparative method reviewed.
Oxford: Oxford University Press, pp. 39 - 71.

Peiros, I., 1989, *Languages and Sociolanguages.*
Paper read in the Institute of Ethnography, Soviet Academy of Sciences, Moscow.

Peiros, I., 1995, *Proto Katuic Comparative Dictionary.*
In Pacific Linguistics C-132. vi+192 pp.

Peiros, I., 1996, *The Vietnamese etymological dictionary and 'new' language families.*
In Pan-Asiatic linguistics. Mahidol University: Salaya, vol 3 pp. 883-895.

Peiros, I., 1998, *Linguistic Prehistory of Southeast Asia* (396 pp.).
In Pacific Linguistics, scheduled for publication in 1998.

Peiros, I., MS *Proto Palaung-Wa reconstruction and Comparative dictionary.*

Peiros, I., and S.Starostin, 1996, *A Comparative Vocabulary of Five Sino-Tibetan languages.* Fasc.1-6.
Melbourne: Department of Linguistics and Applied Linguistics.

Peiros, I., and S. Starostin, in progress, *Lexicostatistics revisited.*

Ross, M., 1995, *Some current issues in Austronesian linguistics.*
In Tryon, D., ed., Comparative Austronesian Dictionary. Part 1: 45-120.
Berlin, New York: Mouton de Gruyter.

Simpson, J. *How Warumungu people express new concepts.*
In Language in Central Australia, 4: 12-25.

Starostin, S. 1989, *Comparative Linguistics and Lexicostatistics.*
In Lingvističeskaja reconstruktsija i dreonejšaja istorija Vostoka (Linguistic reconstruction and ancient history of the Orient, in Russian).
Moscow, Nauka, part 1: 3 - 39.

Tryon, D.T., ed., 1995, *Comparative Austronesian Dictionary.* Parts 1-4.
Berlin, New York: Mouton de Gruyter.