1

INTRODUCTION

*Though it turned out wrong, the steady-state theory was a 'good'
theory in that it made very clear-cut and testable predictions; . . .
A 'bad' theory, in this sense, is one that is so flexible that it can be
adjusted to account for any data. The eminent . . . physicist
Wolfgang Pauli would deride such vague ideas as 'not even
wrong'*[1]

Martin Rees

1.1. Aims of this book

It is often claimed (see, for example, Dixon 1997: 28) that the family-tree
model of historical linguistics is realistically only applicable to a small
number of language families, namely Indo-European, Semitic, Polynesian
and Uralic. The purpose of this book is to examine closely the foundation of
this claim for Uralic, by bringing together the main linguistic and non-
linguistic evidence in one volume.

My conclusions on the origin and nature of the Uralic languages will
differ in several important respects from the conventional view. The detailed
analysis which is carried out in this book has uncovered a total absence of
scientific evidence in favour of the notion that the Uralic languages form a
language family, that is, a genetically coherent group of related languages.
Therefore, in short, I shall conclude that Uralic is not a valid node.

1.2. The standard Uralic theory

The languages forming the Uralic family are spoken in north-eastern Europe
and parts of Siberia. They comprise Finnish, Hungarian, Estonian,
Samoyed, Vogul, Ostyak, Lapp, and many other languages. A typical
family tree[2] is shown in Figure 1.1, which is adapted from Austerlitz
(1987: 178), and a map showing the distribution of the Uralic people and
languages is given at the end of the book (from 'The Gallen-Kallela Museum
& Authors & Artists' (eds) (2000)).

In addition to this conventional family tree, many alternative models have
been published to account for the development of the Uralic languages and
their closest relatives. Broadly, these range from models which question the

```
                              *URALIC


         *FINNO-UGRIC                        SAMOYED


   *UGRIC    *FINNIC (NON-UGRIC)   NORTHERN              SOUTHERN
                                   Yurak-Sam. = Nenets   Selkup = Ostyak-Sam.
                                   Yenisei-Sam. = Enets  (Sayan-Samoyed dialects)
                                   Tawgi-Sam. = Nganasan Kamas-Sam. = Kamassian[3]
                                                         Mator-Sam. = Motor

 *OB-UGRIC   Hungarian = magyar
 Ostyak = Khanty
 Vogul = Mansi



 *PERMIAN            *VOLGAIC          BALTO-FINNIC
 Zyrian = Komi       Cheremis = Mari   Livonian          Finnish = Suomi

 Votyak = Udmurt     Mordvin = Mordva  Estonian = Viru   Karelian

              LAPP = SAAMI  LANGUAGES
```
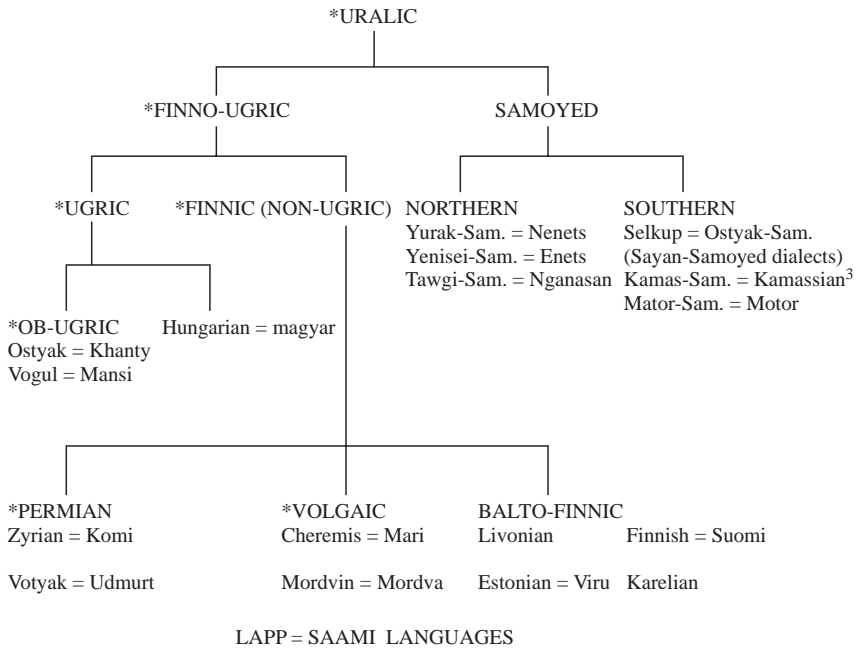
Figure 1.1 The Uralic language family tree as usually reported in textbooks

genetic affinity of the Uralic languages – some of which reject the proposition that they have any genetic relationship at all – to interpretations that go the other way and include other languages, such as Altaic or Yukaghir, in the genetic unity. These various interpretations are reviewed in Chapter 3.

Amongst those scholars who support the mainstream model, and who believe that the Uralic languages form a unique genetic family that is unrelated to any of its neighbours, one finds a range of interpretations regarding the evidence on which this model is founded. For example, some researchers acknowledge that the Uralic corpus is significantly irregular, such as Janhunen (2000: 64):

> the irregularities present in the Uralic etymological corpus are a natural consequence of the early dating. . . . It is therefore no wonder that there is hardly a single Proto-Uralic etymon that would have a perfectly regular representation in every single Uralic idiom. This is, however, not a reason to dismiss the family as a whole.

However, other researchers, in common with textbooks, normally report that the Uralic corpus is essentially regular. For example, Abondolo ((ed.) 1998: 8) says

The historical development of these families has been established, by means of the comparative method, to a degree of precision which is both predictive and productive. 'Predictive' means that given form X in language Y we can predict . . . what the form of its cognate, form Z in related language W, will be. 'Productive' means that such predictions will either prove to be correct, buttressing and fleshing out the detail of the family's genetic integrity, or they will fail, forcing the investigator to re-hone his or her tools and to ask different, perhaps more penetrating questions about the family's internal relations.

Whatever the interpretation, the dominant model remains a simple, clear-cut picture where the Uralic languages form an 'ideal' language family. The following are the basic (sometimes implied or hidden) assumptions, or 'tenets', that are believed by most of the scholars who accept this main-stream approach. For clarity in this book I shall describe the following as forming the 'standard Uralic theory' or the 'textbook interpretation':

Tenet 1: The modern Uralic languages derive from a single genetic parent called Proto-Uralic, and they have no genetic relationship with the surrounding languages. Although it is recognised that there are correlations with languages outside the Uralic group, these are considered to be the result of borrowing or chance resemblance. In particular, the words of Turkic origin present in Hungarian are all classified as borrowed and, for most of them, the exact chronology and donor language has been clearly identified.

Tenet 2: Most of the Proto-Language, and the lower nodes of the family tree, have been reliably reconstructed. The reconstruction extends to all the expected features, including phonological and morpho-logical structure, and is generally regarded to be consistent with the available data.

Tenet 3: Proto-Uralic was spoken in the area of the Ural mountains, at least 6,000–8,000 years ago. The location and antiquity are evidenced by the reconstruction in Proto-Uralic of the names of plants which were present in the area at that time. The antiquity is also supported by the very poor quality of the Uralic corpus and by the fact that Proto-Uralic contains loan-words of Indo-European origin which are very old.

Tenet 4: The reconstructed family tree is consistent with the correlations that are observed in the modern languages. For example, the first binary splitting of Proto-Uralic was very ancient, and it produced Proto-Finno-Ugric and Proto-Samoyed. This splitting can be confirmed by reconstructing shared archaic features. The modern languages under the Finno-Ugric node are comparatively close to

one another, and there is a measurably greater distance from the languages under the Samoyed node.

Tenet 5: The daughter languages which have been formed as a result of subsequent, ordered, binary splits from the mother language contain appropriate shared innovations and retained archaic features consistent with the model.

Tenet 6: The current sound-system of the daughter languages is the result of sound-changes that are essentially regular and systematic.

Tenet 7: The ancient Proto-Language was rich in morphology.

In this book, I intend to demonstrate that these tenets are not supported by the evidence. Whilst there are undeniable correlations between the Uralic languages that can be understood using a genetic model – or using alternative models – these correlations are also present, and just as strong, in some of the surrounding languages. I believe that the above tenets have now become myths that have the effect of holding back research in the field, and that a revision of the entire paradigm is in order.

1.3. Method

1.3.1. *Evidence vs opinion*

My central point of departure from the conventional approach to Uralic studies stems from the fact that the literature frequently fails to differentiate between scientific evidence, which by definition must be falsifiable, and opinions or interpretations, which are usually not falsifiable. As a result, a reader may be forgiven for becoming confused as to exactly what is opinion and what constitutes scientific evidence in the field.

It is important to be clear at this point what is meant by 'scientific evidence'. Much of the literature on the Uralic language family is founded on a network of self-consistent assumptions and reconstructions, and a significant part of this book is dedicated to quantifying and examining them. I do not claim that these reconstructions and assumptions are wrong. They may indeed represent valid opinions or interpretations on the origin of these languages. However, it will be a recurring theme throughout this book that many of them can be demonstrated to be 'not even wrong', that is, there is no objective way to test whether the assertion is true or false. These items must properly be labelled opinions or speculations, since they do not constitute scientific evidence on the criteria normally adopted in scientific research.

For example, in Chapter 4 we observe that the accepted reconstruction of the P-Uralic node contains more sound-rules than regular etymologies. If one regards the reconstruction as a 'scientific model', its significance would be discounted because it contains more 'adjustable parameters' (that is,

sound-rules) than items of evidence. The reconstruction is therefore very flexible in fitting the data. In rejecting theories based on models of this sort, Rees (1999/2000: 76) says:

> A 'bad' theory . . . is one that is so flexible that it can be adjusted to account for any data. The eminent, and arrogant, physicist Wolfgang Pauli would deride such vague ideas as 'not even wrong'

These observations mean that the reconstruction must be viewed as an exercise in stating the sound-laws of the P-Uralic node, if the node exists. The reconstruction is a consequence of, and not evidence for, the existence and validity of the P-Uralic node. This appears to accord with the method and purpose of the original author. Rather than attempting to collate evidence in support of the P-Uralic node, its existence and validity are assumed throughout. The purpose (Janhunen (1981a: 246), my translation) is to give 'at least a general picture of the sound-structure of the Proto-Language', whilst stressing that important elements of the reconstruction are 'totally tentative' and that there is 'the possibility of alternative interpretations' (see Chapter 4 for details).

The question of the status of this P-Uralic corpus merits a 'second opinion' in the form of an alternative test. In Chapter 5 we compare the same corpus with a 'control case', equivalent to random words that are matched using statistically equivalent criteria. We report that these two samples – the true Lexicon and the random words – are statistically almost identical. In other words, the test fails to distinguish whether the corpus is the result of a true linguistic connection or chance resemblances. To take an analogy, if the Uralic corpus were a medical drug, it would be rejected on the grounds that its performance is not discernibly different from a control-case placebo.

These observations do not demonstrate that the reconstructions are wrong. It is certainly a valid opinion that at least some of them represent a true linguistic correlation. I hold this opinion, at least for a small number of these lexical items. But they do not constitute scientific evidence because, both according to conventional methods and according to more detailed statistical methods, they cannot be distinguished from accidental look-alikes.

To take another example, the P-Uralic corpus is (often) recognised to be small and to contain many irregularities, contrary to what would be expected for a language family of this sort. One possible explanation is that the Uralic languages are not related in the way that is usually stated. However, the accepted explanation is that the languages are genetically related but the language family is very old, so that the lack of quality and quantity is a consequence of the great time depth. In order to distinguish between these alternative explanations, independent evidence is required. As we shall see in Chapter 7, there is in fact no independent evidence that the

family is very old. The only evidence is – in a circular argument – the very lack of quality and quantity in the Uralic corpus itself. In fact, the antiquity of the P-Uralic node is a consequence of, and not evidence for, the assumption that the U languages are genetically related.

A third example is the way in which proper names are interpreted in Uralic studies. As we shall see in Chapter 2, the concept of the Uralic theory originated in an apparent similarity of proper names (*hungarus*-*Yugria* and *magyar*-*Mansi*). This latter ethnonym is discussed at length in modern textbooks, where it is cited as evidence in favour of the Uralic theory. This approach is perfectly respectable in principle, since proper names are sometimes regarded outside Uralic studies as particularly stable forms of language. For example, Renfrew (1987: 21) says

> place names . . . do continue to be used in a given area long after their original meaning has been forgotten, so that words belonging to a much earlier and pre-literate form of language can be preserved by spoken tradition, and first set down in writing long after the original language form has disappeared.

However, as we shall see in Chapter 9, the connection *hungarus*-*Yugria* is no longer regarded as valid in most modern textbooks, and the connection *magyar*-*Mansi* is made through a reconstruction that differs from the historically attested form and is linguistically ad hoc. Further, there is a significant number of other connections between proper names in the same geographical region, which span the supposed boundary of the Uralic area, and which therefore are inconsistent with the Uralic theory. Some of these parallels have clear characteristics that mean they cannot be the result of chance resemblances. This evidence is well known in Uralic studies, and it has become common practice to 'explain it away' by assuming that one can deduce nothing from proper names, because they are highly susceptible to change (with the exception, of course, of *magyar*-*Mansi*).

The etymology *magyar*-*Mansi*, its assumed significance, and the practice of disregarding other proper names so as to avoid their contradiction with the Uralic theory, might, for all we know, all be correct; but they are ad hoc: they are not even wrong, because there is no objective way to tell if they are true or false.

Textbook writers and my colleagues in Uralic studies are certainly aware of the distinction between evidence and opinion, and it is not due to sloppiness that this distinction is so often blurred. There is a much deeper assumption at work here. It is believed that the question of the origin of the Uralic languages was settled over 100 years ago by the founder of the field, a German researcher working in Hungary named Budenz. If you assume that the Uralic language family was scientifically established beyond doubt over 100 years ago, then the assumptions referred to above cease to be ad hoc: they are motivated by the established model or paradigm.

This is why my treatise must begin by describing the historical foundation of the Uralic paradigm, and in particular the chief question posed by Budenz, namely the relationship between Hungarian, Turkic and the languages of north-eastern Europe. As will become evident, the claim that the Uralic language family was scientifically established beyond doubt (over 100 years ago) is, in fact, not correct.

1.3.2. *The history of the Uralic theory*

As we shall see in Chapter 2, the concept of the Uralic language family originated, between the sixteenth and eighteenth centuries, in an apparent similarity between the Latin ethnonym for the Hungarians, *hungarus*, and the Russian toponym for a location near the Ural mountains, *Yugria* – hence the terms 'Uralic' and 'Ugric'.

In the late 1800s, a German linguist working in Hungary, J. Budenz, tried to apply the 'Comparative Method' in order to prove the relationship between Hungarian and the other Ugric languages (which he extended to include other languages of north-eastern Europe, such as Finnish). One can observe today that Budenz did not apply the Comparative Method properly in this corpus; for example, he did not state the sound-rules on which his 'correspondences' were supposed to be based, and arbitrary stretches of meaning were often used. In fact, in examining a sample of Budenz' 'correspondences', 81% are no longer considered valid in the modern literature. Further, as we shall see in Chapter 4, it is now widely accepted that Hungarian is radically different in Morphology, Phonology, Lexicon and Syntax from the other Ugric languages, so that it has proven impossible to reconstruct the Ugric node from the primary linguistic evidence. It should not be surprising to find such a large variance between Budenz' comparative corpus and the modern literature on the same topic, since the Comparative Method was in its infancy at the time; and, indeed, a similar situation is found in the Indo-European context, where, as Fox (1995: 11) says:

> the forms of reconstructed Proto-Indo-European have changed out of all recognition as successive generations of scholars have refined and amended their predecessors' work

More seriously, it usually goes unmentioned in the literature that Budenz' conception of the Uralic family differs significantly from the modern theory he is supposed to have 'scientifically established'. In the modern theory, the Uralic languages are held to be completely unrelated to the Altaic languages. One consequence of this is that all the words of Turkic origin present in Hungarian are now classified as loan-words, and, as we shall see in Chapter 2, much of the study of Hungarian prehistory is based, in the sense described above, on this cornerstone. Yet Budenz made no such claim. Quite the contrary. He classified Hungarian as belonging to the extended

Ugric branch (including the languages of north-eastern Europe) and, in his model, the Ugric branch, the Turkic branch and other Asiatic branches all belonged to the 'wide Altaic family'. He supported this model by identifying specific correspondences that he claimed were indicative of a genetic relationship (from a higher level) between Turkic and Hungarian.

At some point in the development of the paradigm, the Uralic languages came to be 'split off' from the languages left behind in the Altaic family. It is an astounding omission that I can find no original work to substantiate this assumed splitting-off. Nonetheless, in spite of the lack of any original reference, it is generally believed that the uniqueness of these language groups has been scientifically established beyond doubt, yet again through the method of comparative linguistics.

1.3.3. *The Comparative Method*

In the method of comparative linguistics the concept of 'regular' sound-laws is central. It is assumed that one can distinguish genetic relations by the regularity of their associated sound-correspondences and sound-changes, whilst words that do not obey regular sound-laws are classified as merely borrowed or accidental look-alikes. The cumulative effect of many words, all obeying the same sound-laws regularly, is supposed to establish the correlations to a high degree of probability.

Unfortunately, all of these elements are untestable or completely missing in the Uralic context. We have already identified above that the accepted reconstruction of P-Uralic contains more sound-rules than regular etymologies, so that there is no 'cumulative effect' in this corpus. Most other research on the Uralic family is silent on what the sound-rules are actually supposed to be, or at best only a handful of them are identified. For example, the main Uralic dictionary, *Uralisches Etymologisches Wörterbuch* (UEW), does not state the phonological criteria on the basis of which its 472 P-Uralic etymologies have been established, and it is not easy to infer them from the etymologies that are listed because of inconsistencies in sound and meaning. This means it is impossible to apply the Comparative Method to any testable degree using this evidence.

Turning to the way in which linguistic items are classified, there is an important bias in the way the Comparative Method is usually applied in the Uralic context. This bias is most evident when one examines the relationships between the Uralic languages and the apparently similar nearby languages to the East.

The method proceeds as follows. One first identifies the languages that are supposed to be genetically related. From these languages, one then reconstructs the relevant Proto-Forms, Proto-Phonemes and sound-changes (as we have noted, these are in fact usually only partially stated in publications). Words that match the predictions of the model can now be cited as

'evidence' in favour of the original assumptions, whilst words that do not match the predictions are dismissed as, variously, in error, irrelevant, chance similarities, *Wanderwörter*, borrowed or (for particularly good matches that cross the Uralic boundary) due to 'convergent development' or 'regular direct borrowing'. Alternatively, as Abondolo ((ed.) 1998: 8) suggests, words that do not match the predictions force the researcher to 're-hone his or her tools and to ask different, perhaps more penetrating questions about the family's internal relations', presumably as an alternative to acknowledging them, more simply, as items of evidence that do not fit the model.

Even with the die biased in favour of the Uralic language family in this way, we shall see in Chapter 6 that many words match better with languages outside the Uralic family than within the family. For example, Sinor (1975: 251) says:

> I find it hard to believe that a correspondence as **flawless** in form and meaning as that existing between the Uralic and Tunguz forms could be coincidental . . . I would not like to exclude the possibility that the . . . words . . . are **direct borrowings** from Tunguz (bold is mine)

In a situation where several researchers recognise within the field that the Uralic corpus is very small and contains many irregularities, why are these 'flawless' correspondences not interpreted as evidence of a genetic relationship? What distinguishes a 'borrowing', which can be identified because it is 'irregular', from a 'direct borrowing' which is the opposite? The only justification for these classifications seems to be that there is an a priori assumption – which is held to have been settled historically – about which languages are genetically related and which are not. At this point one may feel entitled to ask how the literature would have reported this evidence if, by an accident of history, the Uralic paradigm had developed in a different way. It is hard to believe that the 'flawless' correspondences reported above would not have been cited as strong evidence for a genetic relationship between Tungus and some of the northern Uralic languages. Conversely, it is hard to believe that the acknowledged very poor correspondences between (say) Hungarian and the other Ugric languages could have failed to be dismissed as in error or not significant.

A similar bias has also been noted outside the Uralic sphere. For example, as Fox (1995: 63) says, the fundamental problem lies in:

> assuming that forms are cognate because they can be reconstructed with the same proto-phoneme, where the proto-phoneme is itself the result of assuming that they *are* cognate.

In addition to the problems in the way the Comparative Method is usually applied, there are now acknowledged to be inconsistencies between the evidence from other language groups and the founding principles of the method itself. The fundamental assumption of the Comparative Method is

the 'regularity principle', according to which inherited words obey regular sound-laws whilst borrowed words are irregular. Recent studies in linguistics have demonstrated that these foundations are not supported by the evidence.

As we shall see in Chapter 6, borrowed words tend to become assimilated into the sound-structure of the recipient language rather rapidly, so that they become indistinguishable from inherited words after only a few generations, if not at the very time of borrowing. This means that, in the absence of historical records or other types of information, it is difficult, if not impossible, to distinguish borrowed words from inherited words linguistically.

Secondly, there is evidence that the majority of sound-changes do not in fact proceed regularly after all, because of the disturbing interference of several factors, which include social, geographical and generational variation; mixture of languages and dialects due to contact; and the intricate, often arbitrary processes by which changes diffuse, complete or abort. As Lass (1993: 179) says:

> It is really only within the past two decades or so that Neogrammarian *regelrecht* correspondences have come to be naturally interpretable as the end results of long-term processes of much messier change, involving variation, and often lexeme-specific mutation

This counter-evidence has arisen from several areas of linguistic studies: 'languages in contact' (since Weinreich 1953), Dialectology (see, for example, Chambers & Trudgill 1980: 37–38), Sociolinguistics (see, for example, Labov 1963, 1972, 1980, 1994; Weinreich, Labov & Herzog 1968) and from the field of 'lexical diffusion' (see, for example, Wang 1969, 1979; Chen 1972). For the question of the 'regularity' vs 'non-regularity' of sound-changes see Schuchardt (1885), Labov (1981) and Campbell (1996). For a general review of the strengths and weaknesses of the Comparative Method see Fox (1995) and Lass (1997).

1.3.4. *The programme followed in this book*

In this book I shall try to describe, as clearly as I am able, the evidence, the assumptions and the models that are used to account for the Uralic language family. I shall review both the mainstream 'textbook' interpretation and the many alternatives that appear in the literature. I shall refer to primary evidence wherever possible in order to place these interpretations into the appropriate context. I shall always try to separate out the scientific evidence from the assumptions, the interpretations and the speculations.

To help in this separation, I shall use statistical methods as well as conventional linguistic methods of analysis. I shall report data from the fields of history, Phonology, Lexicon, Morphology, Morpho-syntax, Typo-

logy, Onomastics and, briefly, other non-linguistic areas such as archaeology and genetics.

Although I believe the use of the Comparative Method can be misleading in this context, for the reasons outlined above, I shall, where appropriate, review the interpretations that use this method, as dispassionately and as positively as possible, in their own terms – that is, assuming that the Comparative Method is applicable and genuinely applying it together with its known advantages and limitations.

At this point I must ask for the indulgence of my colleagues and readers. As identified above, the current approach to Uralic studies is based on an interlocking network of self-consistent assumptions, interpretations and reconstructions. As we shall see in Chapter 2, re-interpretations of historical evidence, so that it appears to be more consistent with the mainstream model, are also involved. Each of these assumptions, reconstructions and interpretations reinforces another, in a kind of 'Gordian knot'. In my attempt to disentangle some of the more 'Gordian' aspects, it is inevitable that I shall use evidence from sources in support of an interpretation with which the original authors might not agree. Therefore, despite my best intentions, there is always the danger that I might misrepresent the views and interpretations of the scholars who deal with these topics. If this is the case then I proffer my apologies in advance.

In particular, I shall use many examples from Rédei's UEW (1986–91), because this goes back systematically and comprehensively to the original sources, reporting the actual languages and dialects both within the Uralic area and to some extent outside it. I shall refer extensively to Abondolo ((ed.) 1998), which is the most recent English-language publication with a detailed description of all the Uralic languages, and which also largely mirrors the conventional interpretation. I shall examine in detail the reconstructions of P-Uralic and of the rest of the family tree, by Janhunen (1981a) and Sammallahti (1979, 1988) because these represent closely researched scientific models with clear criteria that are amenable to quantitative scrutiny. Regarding the specific subject of the parallels with languages outside the Uralic area, I shall pay particular attention to the comprehensive works of Sinor (1975, 1988, 1990), Ligeti (1986) and Johanson & Csató ((ed.) 1998).

Although my interpretation of the evidence reported in these sources will differ in many respects from those of the original authors, these sources stand by themselves as sources of original data and they are at the foundation of the field, whatever interpretation is chosen by the reader.

No review of this nature would be complete without my personal interpretation of the relationships between the languages involved and the processes that led to them. This is given at the appropriate places. Here, I want to clarify that I am not a proponent of any specific linguistic 'school'. My personal view is that there are several processes that can equally well

explain the correlations that are observed today amongst these languages; these include genetic inheritance, areal convergence, and Renfrew's (1987) model of the cumulative effect of small random movements of individuals. All of these processes have their place, and it is my central claim that it is usually not possible to tell them apart. Sometimes, in explaining my personal interpretations, I shall select one or other of these processes for the sake of concreteness; this is not intended to exclude the other processes unless specifically stated.

It is possible that Uralic studies are not alone in being susceptible to re-examination; however, the purpose of this book is to examine the Uralic languages and I shall remain true to this focus. I speculate on any applicability outside the Uralic sphere in the final chapter.

1.4. SYSTEMATIC RE-INTERPRETATION OF EVIDENCE COUNTER TO THE STANDARD THEORY

In this book I hope to show, by careful examination of the evidence and the literature, that the tenets of the Uralic theory are not supported by the evidence, and indeed there is a large body of evidence counter to them. However, it is striking to observe in the literature how the counter-evidence becomes consistently diluted as it makes its way in turn from reported data, to authors' conclusions, to the general literature and to English-language textbooks. In original papers, the individual items of counter-evidence are usually recognised explicitly. However, when drawing conclusions, authors usually say that they neglect them, or that they consider them not relevant, or in other ways minimise their importance. This process appears to be repeated at each step in the chain, so that, in the final step, textbooks usually present a highly idealised picture. The result for the field is that the counter-evidence is systematically minimised or re-interpreted, and the evidence, as propagated into the general literature, seems to be consistent with the conventional paradigm.

As was observed by Kuhn (1970), it is quite normal in science for evidence that is counter to an established paradigm to be consistently minimised or re-interpreted. Typically the individual items of counter-evidence are not strong enough on their own to mount a challenge to the established paradigm. After all, any individual items of evidence may be subject to misinterpretation or statistical error or other problems. The result is that authors will often willingly choose a presentation and interpretation that makes the work appear as consistent as possible with the established theory. This phenomenon is sometimes extended into the peer-review process, so that if a paper submitted for publication describes such evidence and reaches a conclusion that is inconsistent with the accepted paradigm, it will probably

(and arguably correctly) be rejected for publication unless the mismatch with the paradigm is minimised.

Here are some examples of this process at work in the Uralic context.

A linguistic model that is established in other language families may be 'adjusted' in so far as it applies to Uralic studies if it is inconsistent with the Uralic evidence. In this way the evidence is made to appear consistent with the Uralic theory. For example, some of the basic words at the heart of the Comparative Method are kinship terms and number terms. According to the comparative model, these are particularly stable forms of language (see, for example, Comrie ((ed.) 1987b: 24–5) in the Indo-European context). However, most of these words are irregular in Uralic. Rather than acknowledging this as evidence that does not match the established model, it is the model itself which is adjusted specifically for the Uralic context. This irregularity is classified in Uralic as if it were a systematic phenomenon. It is called, in the jargon of the field, the 'typical lability' (Janhunen[4] 1981a) of the 'affective words' (UEW[5]). It is now seemingly part of the model in Uralic studies that these words, which are usually considered at the heart of the Comparative Method, are irregular because of their 'affective character' and 'high frequency of use' – but I can discover no explanation as to why affective character and frequency of use should produce this effect in the Uralic languages but the opposite effect in, for example, Indo-European languages.

It is sometimes the evidence that is 'corrected' (or 're-interpreted') if it is inconsistent with the theory. For example, as we shall see in Chapter 2, the earliest records, from the 9th/10th century, consistently refer to the Hungarians as 'Turks'. This is directly counter to the thesis of a Uralic origin. It is therefore assumed that the contemporary sources had become 'confused': the Hungarians were not Turks, but had become similar in appearance and behaviour to the Turks through a long period of co-habitation. This long co-habitation is supposedly supported by the writings of the Byzantine emperor Constantine Porphyrogenitus, who is usually quoted as stating that the Hungarians were in contact with the Turks for about three hundred years. His actual record states that a group of seven Turkic clans [re-interpreted as the Hungarians] were in contact with the [Turkic] Khazars for three [re-interpreted as three hundred] years, but historians simply assume that the emperor, too, was in error, so that in modern accounts he is usually held to refer to different populations and to a different length of time than is actually stated in the original text. The emperor's 'errors' and the re-interpretations to make the text consistent with the Uralic paradigm may go unmentioned in textbooks or even in specialist literature (see as a recent example Rédei[6] 1998: 57).

In another process of minimisation, counter-evidence may be reported, but its significance may be diluted in conclusions. For example, in numerous publications on various Uralic and Altaic correlations, Sinor reports striking

shared lexical and morphological features. We have already noted above the 'flawless' correspondences that the author has identified between Tungusic and Uralic, and these are reaffirmed in a later paper (Sinor 1988: 738–9; see discussion in Chapter 6)

> I am quite certain that if from all the Uralic and Altaic languages only the [Altaic] Northern Tunguz and [Uralic] Ob-Ugric were known, no one would deny their genetic relationship.

This is directly counter to the predictions of the accepted models, and perhaps in other scientific spheres the reader would expect an explicit statement to this effect. However this is not the case. Instead, the paper appears to reject the concept that the accepted models might be falsifiable at all:

> Uralic, Altaic, and Uralo-Altaic comparative linguistics should shake themselves free from simplistic – black and white, yes and no – solutions

Finally, this paper, which contains strong counter-evidence to the accepted models, concludes with merely a diluted call for further research leading to a possible 'revision of the traditional approach to Uralic-Altaic comparative studies'. Its final words are:

> I can but conclude with a suggestion I made years ago, that Uralic and Altaic comparative linguistics should not ignore each other, and that the truth should not be proclaimed but searched for.

Finally, where conclusions that do not fully support the conventional paradigm are clearly stated in original papers, these conclusions may be diluted in works that quote them or are based on them. For example, as we shall see in Chapter 4, Sammallahti (1988) reconstructs the full conventional Uralic family tree, starting from Janhunen's (1981a) reconstruction of the P-Uralic node. About 50% of the examples that are quoted in support of the reconstructed family tree are in fact identified as 'irregular', 'problematic' or 'troublesome' in the original work of Janhunen. This is not mentioned. On the contrary; for example, it is asserted (1988: 479) that Janhunen's corpus contains '140 regular cases' when the original paper classifies only 94 cases as acceptable parallels, some of which are identified to have irregularities that the author believes can plausibly be explained. A reader who has not read this original work of Janhunen (in Finnish) may be led to the false conclusion that there is a great deal of regularity and consistency in both corpuses. This picture becomes even further idealised in textbooks; for example, Csepregi (1998) gives selected examples showing the apparent complete regularity of a few reflexes, omitting mention of any recognised irregularities at all. We shall see further details in Chapter 4.

These are not isolated examples. I believe that a careful examination of the counter-evidence reveals a wide scope and a clear pattern: the counter-evidence has now reached the point of critical mass. A revision of the

paradigm is now in order. The time has come for all the evidence and counter-evidence relating to the Uralic languages to be brought together in a systematic and comprehensive way.

1.5. KEY PATTERNS IN THE EVIDENCE

My personal interpretations regarding the methods that are conventionally used to classify correlations between languages, and specifically regarding the nature of the correlations between the Uralic languages, is given at the end of the book. Here, I limit myself to outlining the key patterns to which I believe the reader will objectively be drawn, after examining the evidence set forth in this book.

The first conclusion is that there are a great many structural (morpho-syntactic and typological) elements, and a handful of non-structural elements, that span most of the Uralic languages and which are unlikely to be the result of chance resemblances. Almost all of these common elements are also shared with the other languages that are conventionally classified as Altaic and Yukaghir. The small number of non-structural elements include: some phonological features (such as richness of vowels, paucity of consonants, and some shared sounds and sound-changes), a handful of basic words, and a few basic formants consisting of the most natural, common phonemes.

Secondly, clusters of isoglosses can be identified at several levels of language (Morphology, Phonology, Lexicon, Onomastics etc.) that are broadly consistent with one another. A number of these clusters run right across the conventional boundary of the Uralic area.

Thirdly, with regard to Morphology, the modern Uralic languages have varying degrees of complexity, ranging from the simplest, such as Vogul and Ostyak, to the most complex, such as Finnish, Hungarian and Mordvin. In the morphologically rich languages, the basic processes through which Morphology is formed (Exaptation and Grammaticalisation) are clearly observable, and one is drawn to the conclusion that the complexity in this area is a recent innovation. Whatever the nature of the ancestor or ancestors, the evidence suggests they were morphologically simple.

Fourthly, whilst researchers in the past have relied mainly on the family tree model to interpret and explain the relationships among the Uralic languages, this model is simply not supported by the evidence. If one is to adopt standard scientific principles, one must conclude that there is no evidence that the Uralic languages form a coherent family in the sense conventionally under-stood by the term, so that Uralic must not only be removed from the list of 'well-behaved' language families, it must also be removed from the list of language families altogether. Instead, the linguistic relations among the languages are best described in terms of intersecting isoglosses.

Finally, the U languages and the nearby languages to the East form a dialectal continuum whose formation may be the effect of several processes, including genetic inheritance and areal convergence, but, on the basis of current evidence it is, in the main, not possible to tell these processes apart.

1.6. Introduction to the literature supporting and opposing the Uralic theory

The U paradigm/theory has been successfully passed on, to this day, through a vast body of articles and books. Amongst the most recent of the publications that propagate the textbook U theory one can include the volume *The Uralic Languages*, edited by D. Abondolo (1998); the paper by L. Campbell (1997): *On the linguistic prehistory of Finno-Ugric*; the book *Finnugor kalauz*, the 'Finno-Ugric guide', edited by M. Csepregi (1998); the book by G. Bereczki: *Fondamenti di Linguistica Ugro Finnica* (1998); the book by K. Rédei (1998): *Őstörténetünk kérdései* 'The questions of our pre-history'; the book by J. Laakso (1999b) *Karhunkieli*, literally 'the language of the bear'.[7] Compare also Esa Itkonen (1998 and 1999), supporting the validity of the family tree model in general and within U studies.

Despite its popularity among linguists and the general public, and despite the fact that it is generally presented as proven beyond doubt, the standard Uralic theory is far from unquestioned. As we shall see in detail in Chapter 3, one can find several different analyses and points of view about the origin and classification of the Uralic language family, which seldom get the chance to be brought to the attention of the broader circles of readers, and therefore to be seriously considered.

Such a contrast between the simple textbook picture, and the, sometimes, contradictory interpretations found in the specialist literature might seem at first surprising. However, as pointed out by Lass (1997: 5–6) with regard to the histories of languages in general:

The histories of languages (as objects available to or made by linguists) are, like all histories, myths. We do have documents for portions of many of our histories; but even these are subject, like scripture, to exegesis: we don't *know* what they mean (the less, the older they are). We do however tell (and believe) stories about them, not just the documents but the languages they supposedly reflect, the reality (phonic, grammatical, semantic etc.) . . . But a venerable rational mythology, apparently grounded in argument and extrapolation from putative evidence, can pose a major problem: the longer it exists, the less succeeding generations or practitioners tend to know, or remember (if they ever knew) or even care about how it came into being, or what supports its main tenets. . . . in many important cases, we may be passing on, as precious and firmly held

beliefs, replicas of assertions that someone somewhere once made, transformed into Articles of Faith

In more detail, within U studies one finds, for example, that documents, archaeological and ethnographic records are scanty in relation to the U languages and people. Publications often refer to the 'mystery' surrounding these languages and the archaeology and ethno-history associated with them. There is clearly a requirement to interpolate heavily between the available data, and this might well increase the potential for a 'mythological' component in putting together the pieces of the puzzle.

There is a body of research recognising the inadequacy of the conventional family tree in accounting for the history and nature of the U languages, which has been often ignored in both the specialist and the general literature. The criticisms raised against (at least some of) the tenets of the standard U theory have recently intensified, particularly by some Finnish, Estonian and Hungarian scholars. These scholars also recognise that the field has to renew itself by taking into consideration, in addition to the conventional concepts, new ideas and methods, including the results of disciplines outside linguistics, mainly human genetics and archaeology. Compare, for example, some of the papers contained in the volumes edited by Julku & Äärelä (1997), by Fogelberg (1999) and by Künnap (2000). Compare also some recent books (such as Kiszely[8] 1992), which challenge the U origin of the Hungarian language and people. In some cases, new models have also been proposed. For example, in one proposal, P-U can been regarded as the lingua franca of the hunters and gatherers of north-eastern Europe. The Balto-Finnic languages are considered as 'mixed' languages, intermingled with Indo-European languages. For Hungarian, the once popular thesis of their Turkic origin has been recently reconsidered.

As pointed out in Section 1.3.3 above, modern linguistic theories have to some extent overtaken the family tree model in general. The academic acceptance of the Uralic theory into the establishment dates back to the end of the 19th century. At that time, the family tree model was practically the only one fashionable and used (the alternative 'wave model', proposed by Schmidt in 1872, was not considered a mainstream model). Since that time, the Uralic theory has been 'passed on' from one generation of practitioners to the next practically as an 'Article of Faith' (as also remarked by Häkkinen[9] 1996: 52). As a result, there has been a very surprising lack of modification of the paradigm, in the face of the deeper linguistic knowledge available today.

Other researchers have independently come to the conclusion that there are a number of myths connected with the U theory, although their views about the nature and classification of the U languages may differ from mine. For example, in a recent review article, *Facts and myths about Uralic studies*,

Salminen (1997b: 86f.) describes what he calls 'the most persistent myths in the field':

> In practically all textbooks, the standard claim is that the Uralic family is a union of two very distantly related groups of languages, called Finno-Ugrian and Samoyed. This claim is false. While it is true that the Samoyed branch is very independent, there is little or no evidence for the rest of Uralic as a unit in its own right

He goes on to say:

> The standard classification continues to split the 'main' branch, i.e. Finno-Ugrian, into Finno-Permian and Ugrian, Finno-Permian further into Finno-Volgaic and Permian, . . . This practice is also unfounded, and originally based on a nationalistic Finnish view which wanted to see the Finnish language literally as the highest spring of the sacred family-tree

Künnap (1997a: 65–6) states that:

> The wish of the majority of Uralists is to take the origin of possibly numerous features of modern Uralic languages back to a single and unitary source – to Proto-Uralic – ; it is psychologically understandable: a simple starting position emerges. Unfortunately it has become clear to date that alongside of this an inadmissible simplification takes place. The simplification which casts aside the abundance and variety of languages, their irregularities and internal contacts. . . . In most cases the reconstruction of Proto-Uralic is declared to be an indispensable methodological mean in the research of historical linguistics. Is it really so if we bear in mind the number of misinterpretations it can create about modern Uralic languages? I am firmly convinced that the methodological mean does more harm than good. A linguistic game with combined rules has been created but its incompatibility with the evidence of modern Uralic languages . . . today becomes more and more evident by the day. Why should Uralists keep playing this game?

Finally, Nuñez (2000: 60) states that:

> There are a number of biases affecting our perception of Finnish origins . . . Among these is a legacy of strong migrationistic and nationalistic orientations and of old linguistic concepts such as family trees and ancestral homelands. Since the beginning of the century, a series of circular arguments have been incorporated in order to create a self-sustaining closed theoretical model that fed and legitimized itself

In the following chapters I shall go into the details of what I believe to be the facts and the myths relating to the U language family, according to the stated aim of this book.