The Family Tree Model and "Dead Dialects": Eastern Middle Iranian Languages

8. The Family Tree Model and "Dead Dialects": Eastern Middle Iranian Languages

YOSHIDA Yutaka
Kyoto University
Translated by KIKUSAWA Ritsuko
National Museum of Ethnology, Japan
The Graduate University for Advanced Studies, Japan

Abstract

Dixon (1997: 46) writes:

> We have just a few scenarios with lengthy historical records—involving Greek, Indo-Aryan, Hebrew and Egyptian—and these extend 3,000 or at most 4,000 years, a small fraction of the 100,000 or so years that language is thought to have been around.

This list of language groups seems to represent a general view shared by linguists as to the availability of written texts. However, there are some languages that have not received the attention they deserve and the Iranian, or the Irano-Aryan sub-branch of the Indo-Iranian languages is one of them. Although they were always behind the more conspicuous Indo-Aryan group, languages belonging to this group have historical documentation covering no less than 3,000 years, and numerous different modern languages of this group are spoken in the vast area extending from Anatolia to China's western border. Thus, the Iranian languages have a lot to offer for diachronic studies, no less than Italic or other Indo-Aryan branches. I will present selected topics from Iranian historical linguistics, in particular the Eastern Middle Iranian dialects of the Pre-Islamic period, and discuss how they relate to the tree model.

8.1. Introduction

This paper discusses issues related to the family tree model and the study of the history of languages from the perspective of philological studies of Middle Iranian languages. When discussing the genetic relationships among languages, a family tree model is often used. However, drawing an accurate family tree for a group of languages requires accurate knowledge as to how the languages developed. Without such knowledge, a family tree could still be drawn but it wouldn't correctly reflect how languages developed.

To illustrate this, I would like to talk about what a tree model would be like if the development of *yatsuhashi* manufacturers was to be drawn. *Yatsuhashi* is a kind of

Japanese confectionary, one of the specialties of Kyoto. Today, there are *yatsuhashi* manufacturers carrying names such as, Yatsuhashi-Hompo, Honke-Yatsuhashi, or Yatsuhashi-Honten and so on. The Japanese words, *hompo*, *honke*, and *honten* all mean a 'head shop' or 'originator', and one interpretation of this situation would be that one of them is the real originator while the others who are claiming to be, in fact are not. If a family tree is to be drawn based on this interpretation, there would be a single line departing from the supposedly original manufacturer connecting to its successors. The others (which claim to be "real" but are not so) would not belong to the tree. Here, it should be noted that this interpretation comes from the presumption that there was only one original *yatsuhashi* manufacturer and the others all imitated it. However, a different tree could be drawn with a different interpretation.

For *yatsuhashi* to develop into the form we see today, each of these *yatsuhashi* manufacturers must have made some contributions, independently applying some modifications in pre-*yatsuhashi* stages. What we see as *yatsuhashi* today, probably is a sum of such accumulated modifications. In this sense, all the manufacturers can be regarded as the originator of each kind of *yatsuhashi*, which must have been unique at the stage when it was developed and put into the market. For instance, *nama-yatsuhashi* 'raw *yatsuhashi*' is very popular today but it did not exist until several decades ago. Obviously, one of these "originators" innovated it, although it is not known who it was and when the innovation took place. With this fact, the assumption that a family tree could be drawn to reflect the development of *yatsuhashi* manufacturers turns out to be wrong, and the question arises as to how the facts can be identified and what kind of model would be appropriate to represent the identified facts.

The point presented in this paper is that historical linguists have to deal with the same kind of issues when drawing a family tree model representing the history of languages. Unless historical facts about the languages are identified, the history of languages cannot be adequately represented by a family tree. To show this point, the following points will be discussed.

1) How changes involving the development of the varieties attested in ancient texts of a language are determined. This includes how regional variations and their chronologically different stages are identified.

2) How innovations that are shared by a dialect of a language (but not others) and a neighboring language, are related to the application of the family tree model.

3) How useful basic vocabulary lists and other linguistic features are in identifying linguistic unity after divergence.

4) How the relative chronology of sound changes identifies the divergence of Eastern Iranian and Western Iranian.

8.2. Iranian Languages

Iranian languages belong to the Indo-Iranian branch of the Indo-European language family. They are characterized by wide geographical distribution, relatively deep time depth, and the cultural diversity of the speakers. The group includes Old Persian, which

goes back to the 6th century BC and Avestan, a language used in the *Avesta*. The term *Avesta* refers to the primary collection of sacred texts of Zoroastrianism, which were originally orally passed down from generation to generation. A written system with unique characters was invented around the 6th century, and the sacred texts started to be transcribed. The exact date when the Avestan language was spoken is not known, but it is considered to go back to an earlier date than Old Persian. In addition, there are six languages whose existence has been confirmed in documents. They are Middle Persian, Parthian, Bactrian, Sogdian, Choresmian, Khotanese, and Tumshuqese. These belong to "Middle-Iranian," which refers to the languages spoken during the period between the conquest by Alexander the Great in the fourth century BC through to the period when most of the Iranian-speaking world became Islamized. The present author's research focus has primarily been on Sogdian and Bactrian. Map 8-1 shows the geographical distribution of the Iranian languages, and Map 8-2 shows areas where Sogdian and other major Iranian languages were spoken during the Achaemenid time.

As mentioned earlier, it is known that Old Persian was spoken about 2,500 years ago. If we go back to Proto-Indo-Iranian, from which all Iranian languages developed, we can trace further back, at least 3,500 years of history. Old Persian and the language of the *Avesta* share many characteristics with the Old Indian language or Sanskrit and it is not difficult to reconstruct the general form of Proto-Indo-Iranian. The oldest varieties of Old-Indian are considered to date earlier than 1,000 BC and Proto-Indo-Iranian, its parent language, obviously goes back beyond this date. Thus, both modern Indian and Iranian languages can be said to have at least 3,500 years of history.
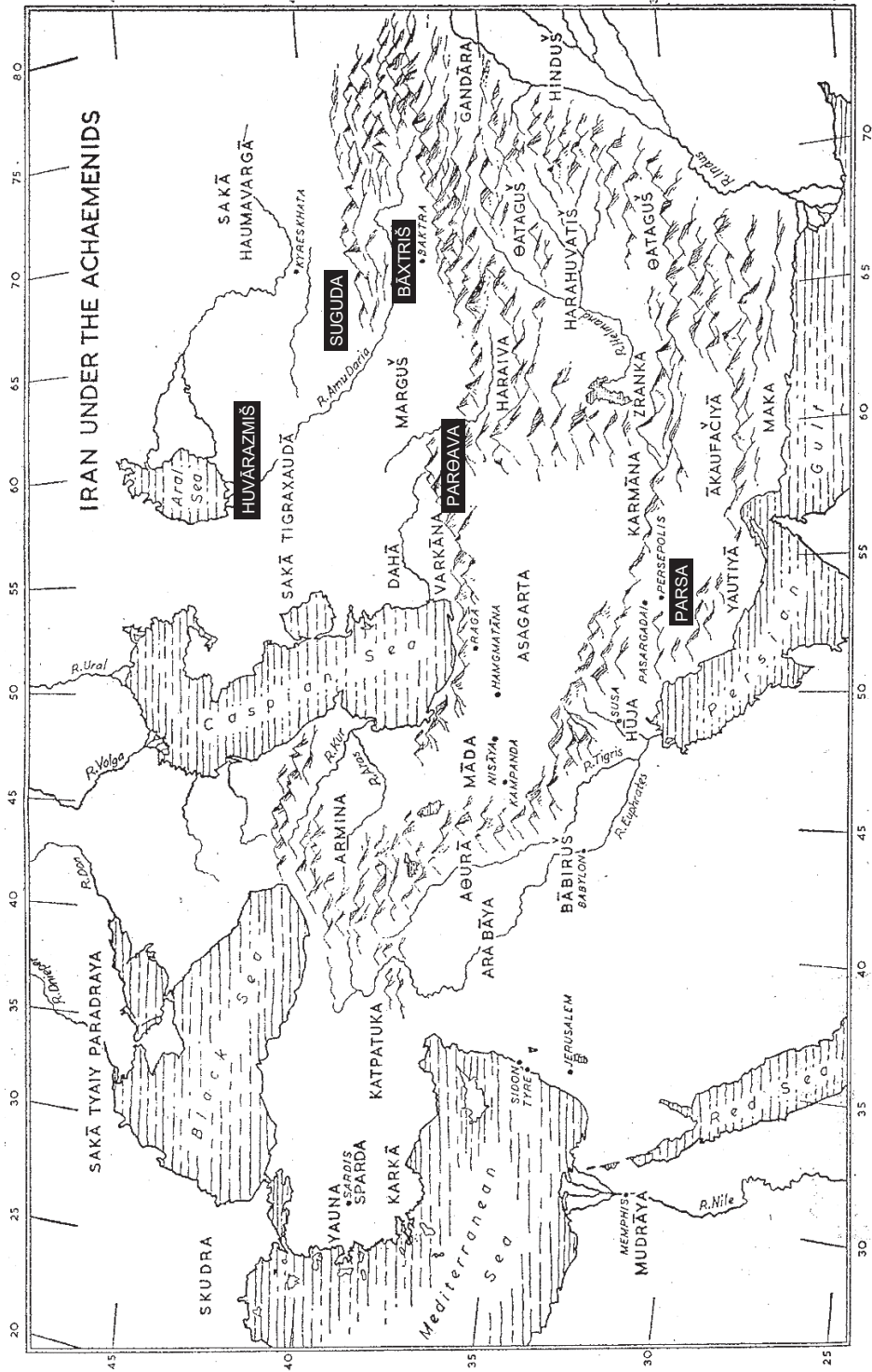
Geographically, Iranian languages are spoken in a wide area. The official language of the Islamic Republic of Iran today is Persian, which has directly descended from Old Persian via Middle Persian. In addition, Iranian languages spread from Asia Minor to Central Asia. These include language isolates, such as Ossetic spoken in the Caucasus. Ossetic is one of the unusual varieties of Iranian, with the speakers' background being Christian while most Modern Iranian speakers are Muslim. The religious backgrounds of the speakers of Iranian languages are diverse. For example, in addition to Zoroastrianism, which is unique to Iranian people, Islam, Christianity, and Judaism were and still are followed. In addition, Buddhist texts in Sogdian and Khotanese have been found and thus it is known that there were also Iranian Buddhists in the eastern areas.

In the field of historical linguistics research on the Iranian languages has been so far rather "low-key", hidden behind Indian languages, which are known for their abundance of written documents. However, I consider that the Iranian languages are ideal for historical linguistic research as well. This is because, in addition to the fact that their long-term linguistic changes are traceable through written records, the languages show a large geographical distribution, and the speakers' cultures are very diverse.

Information about Iranian languages occasionally appears in historical documentation, and based on such information, general changes are often traceable. For example, in the *Shiji* (史記), there is a quotation from Zhang Qian (張騫), who is known for his travel throughout the Iranian-speaking world in the 2nd century BC during the Western Han period. Zhang Qian states that:
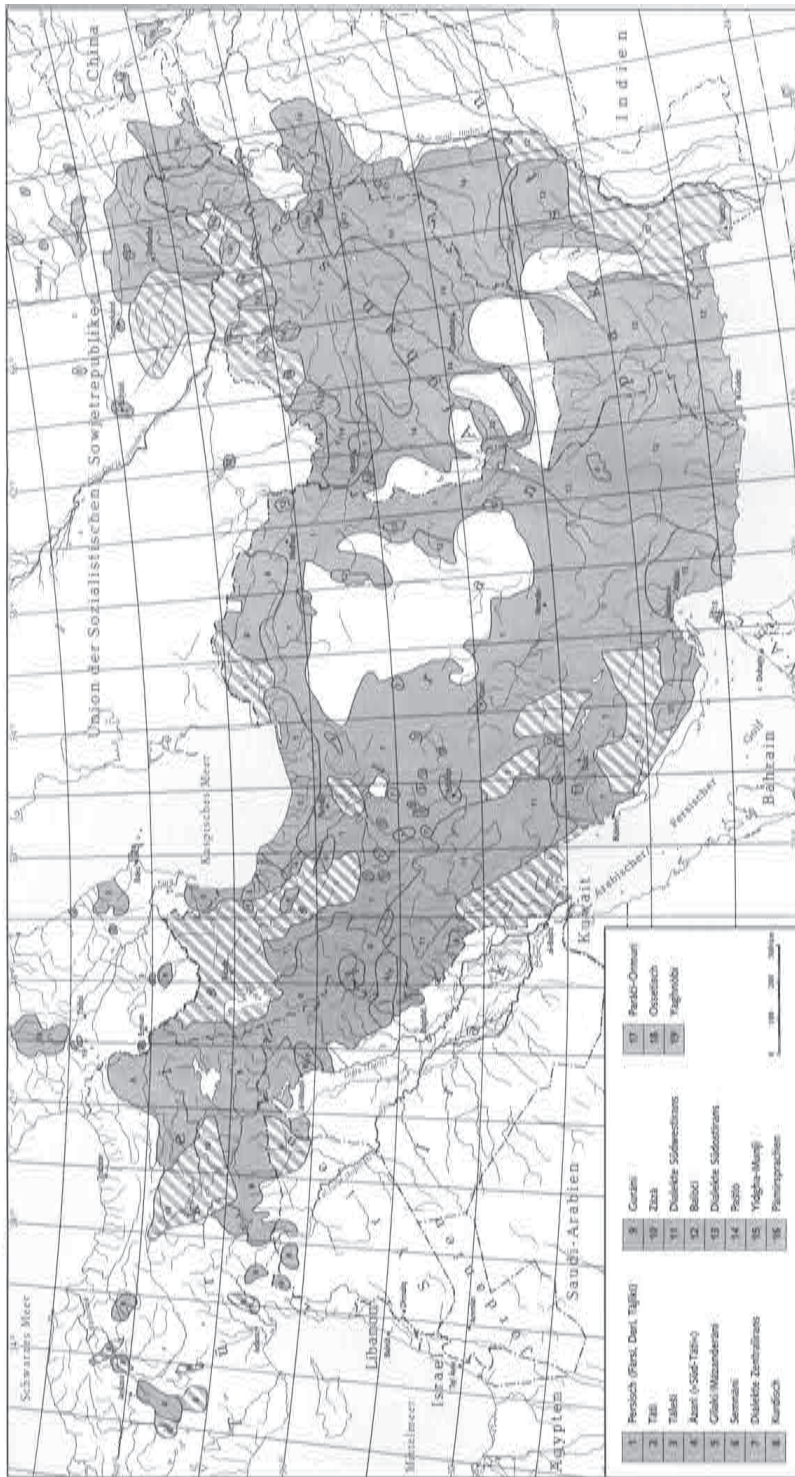
Map 8-1 Geographical distribution of the modern Iranian languages (reproduced from Schmitt 1989)

IRAN UNDER THE ACHAEMENIDS

SAKĀ HAUMAVARGĀ

SAKĀ TIGRAXAUDĀ

HUVĀRAZMIŠ

SUGUDA

BĀXTRIŠ
•BAKTRA

GANDĀRA

HINDUŠ

*KYAESKHATA

R. AmuDaria

MARGUŠ

ΘΑΤΑGUŠ

HARAUVATIŠ

ΘΑΤΑGUŠ

Aral Sea

DAHĀ

PARΘAVA

HARAIVA

ZRANKA

KARMĀNA

ĀKAUFAČIYĀ

MAKA

VARKĀNA

ASAGARTA

PERSEPOLIS•

PARSA

YAUTIYĀ

R. Ural

Caspian Sea

•RAGĀ

•HAGMATĀNA

PASARGADAI•

R. Volga

ARMINA

MĀDA

NISĀYA•

KAMPANDA

SUSA•

HŪJA

R. Tigris

R. Euphrates

R. Don

AΘURĀ

BĀBIRUŠ
BABYLON•

R. Kur
R. Araks

ARA BĀYA

SKUDRA

SAKĀ TYAIY PARADRAYA

R. Dnepr

Black Sea

KATPATUKA

SARDIS•

SPARDA

KARKĀ

YAUNA

JERUSALEM•

SIDON•
TYRE•

Mediterranean Sea

MEMPHIS•

MUDRĀYA

R. Nile

Red Sea

Persian Gulf

R. Indus

R. Haituman

R. Hilmand

Map 8-2   Areas where Sogdian and other major Iranian languages were spoken during the Achaemenid time (from Frye 1984: map 2)

> Although the states from Dayuan west to Anxie speak rather different languages, their customs are generally similar and their languages mutually intelligible. (Sima Qian tr. by Watson 1993: 245)

If we assume that the description in the statement is correct, the interpretation would be that, in the Iranian-speaking world at that time, languages had already diversified to the extent that they were recognized as "different", nevertheless, they were still mutually intelligible.

Talking about "mutual intelligibility," there is an interesting case which implies that Iranian languages were probably regarded as not so different from Indian languages. An Iranian form of the verb *śavati* 'he goes' is quoted in the *Nirukta* about 300 BC by its author Yāska, an Indian grammarian.

> *śavatir gati-karmā kambojeṣv eva bhāṣyate*
> 'the word śavati as a verb of motion is spoken only among the Kambojas.' (Bailey 1971: 64)

The name *Kamboja* seems to refer to a place somewhere on the Iranian side of the border area between India and Iran, which is known as Afghanistan today. The form *śavati* cited in the *Nirukta* is an old form of *šawad* 'he goes', a word used in Modern Persian. Although it is not possible to pronounce on the mutual intelligibility between the Iranian and Indian languages spoken in this area based only on this fact, the wording gives an impression that this Indian grammarian recognized that Old Indian and Old Iranian were not completely different.

In the 7th century, about eight hundred years after Zhang Qian's visit to Iranian speaking areas, a Chinese pilgrim Xuanzang (三藏法師玄奘) visited India via Sogdiana. He states:

> ... the land is called Su-li (= Sogdiana), and the people are called by the same name. The literature (written characters) and the spoken language are likewise so called. (Beal 1906: 26)

A hundred years after Xuanzang's visit, Hyecho (慧超), a Buddhist monk from Silla in the Korean Peninsula, traveled through Central Asia. Regarding the languages spoken in the area, he comments that "The [Sogdian] languages are different from those of other countries" (Yang et al. 1984: 54). This could be interpreted as, by then, mutual intelligibility between the languages had been completely lost. Three hundred years later, there is another report about languages in Sogdiana after Islamization by Al-Muqaddasi. According to his report, the languages had been replaced by Persian. He says that in Bukhara, an oasis city on the Silk Road located in the west of Samarkand, the language spoken there was also Persian. However, in its outskirts, it was not Persian that was spoken but something similar to the one spoken in al-Sughd, the area located between Samarkand and Bukhara. Al-Muqaddasi's report in the 10th century states:

> The language of al-Sughd is unique to it and is approximated by the languages of the rural districts of Bukhārā, which are quite varied, but understood among them. (Collins 1994: 273)

Based on this, in the 10th century, more and more Persian was used in city areas while in the countryside, varieties of Sogdian were still spoken.

Most of the languages in Sogdiana underwent further changes under the influence of several Turkic languages after Persianization. However, in the remotest areas, a unique descendant of Sogdian has survived to this day. In the Yaghnob gorge located deep in the mountains of Tajikistan, which itself is a mountain country, Yaghnobi is spoken. This language is known as a remote descendent of Sogdian and is sometimes referred to as "Modern Sogdian." Incidentally, the word *Yaghnob* etymologically means 'glacier' in Sogdian.

8.3. Linguistic Philology and the Family Tree Model

In this section, I will talk about the stemmas used in philological studies (3.1) and a specific case of the mechanism of micro-level language split that can be identified using text materials (3.2).

8.3.1 Manuscript Classification and Stemma

In philological studies, stemmas are used which resemble a linguistic phylogenic tree. The present author once discovered that one text written in Sogdian script and belonging to the German Turfan Collection (Plate 8-1) in fact does not represent Sogdian but a Middle Chinese poem phonetically transcribed in Sogdian script. He could identify this Chinese text with a hymn popular among the Zen Buddhists of the 10th century Dunhuang entitled *Jingangwuliwen* 金剛五禮文. He then collated thirteen Dunhuang Chinese manuscripts containing the hymn and produced a stemma (Figure 8-1) showing the interrelationship between them.

To draw a stemma like Figure 8-1 one must first produce a so-called "critical text". This is a hypothetical text that is considered to show what the text in different manuscripts could have ultimately originated from. Then, the text in each manuscript is compared to the critical text and matching and non-matching parts are identified. Based on the results, the texts are classified into groups and this classification is represented as a stemma.

Although the stemma appears to show the chronological development of the text, this is not the case. It is a classification based on the similarities and differences that are reflected in the stemma of the texts which are all from the same era, namely the tenth century. In other words, what the stemma shows is a typological and not a historical classification of the manuscripts. If a stemma was to be drawn showing the historical relationships of the manuscripts, what we would need to know is facts about the production or copying processes; namely, when, who and from which text each manuscript was copied.
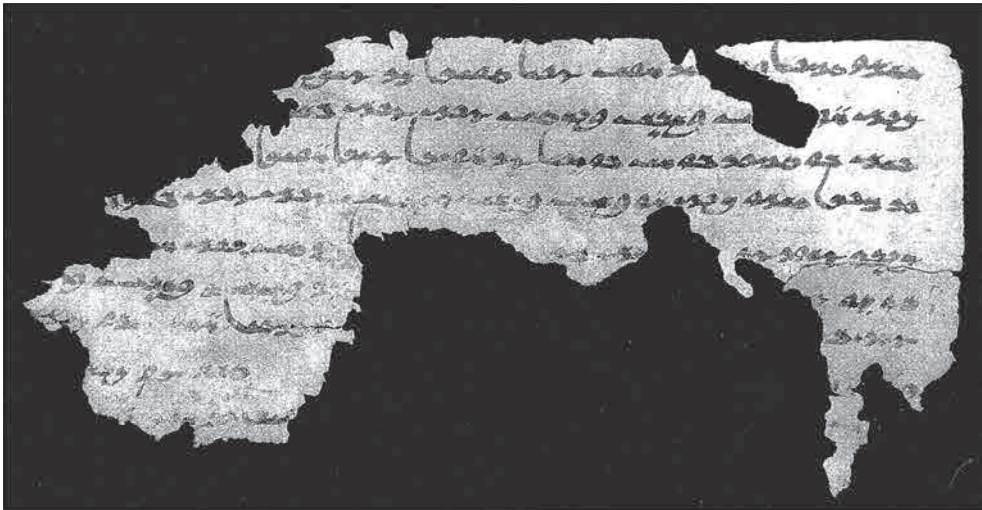
Plate 8-1 Fragment of a Buddhist Chinese text phonetically transcribed in Sogdian script (Mainz 160 + Mainz 627 after Yoshida 1994, p. 360, Depositum der Berlin-Brandenburgischen Akademie er Wissenschaften in der Staatsbibliothek zu Berlin — Preussischer Kulturbesitz Orientabteilung)
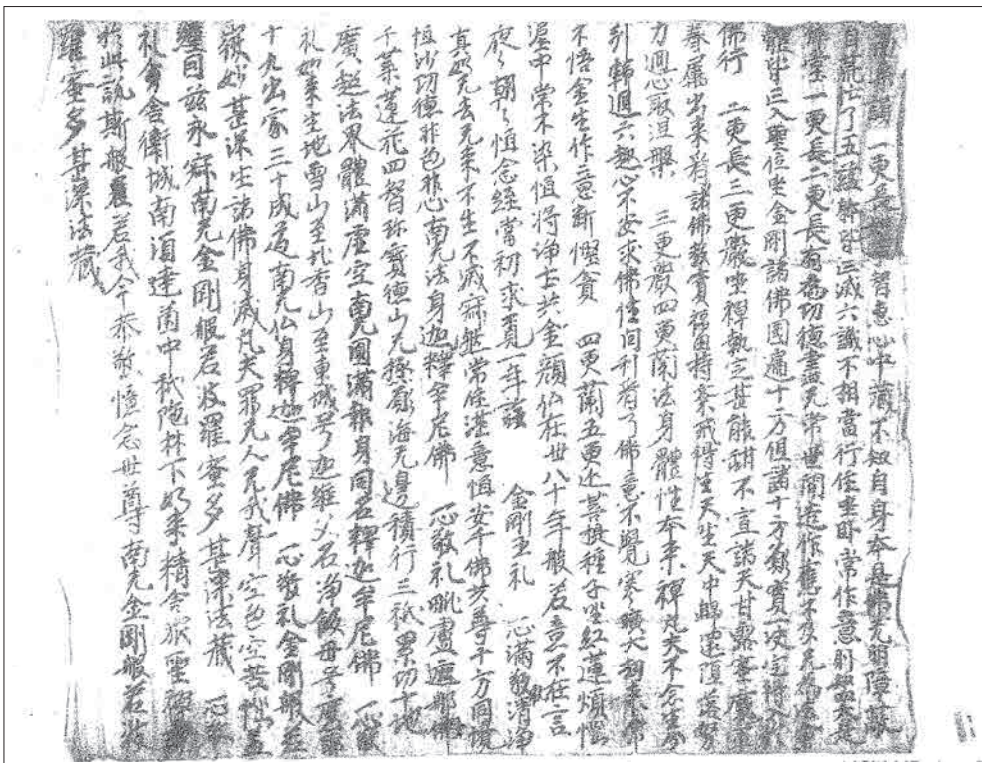
Plate 8-2 Dunhuang Chinese manuscript (The British Library, Or. 8210/S4173) containing the *Jingangwuliwen* 金剛五禮文
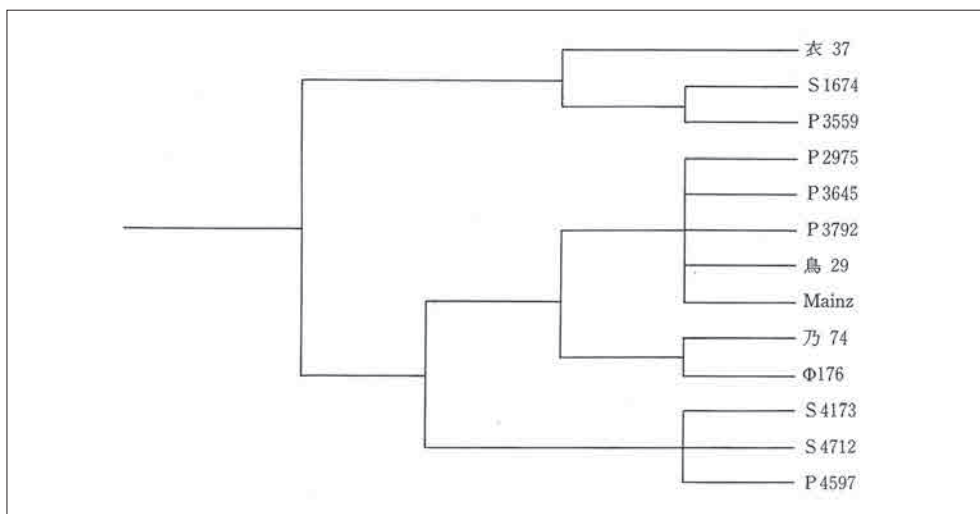
Figure 8-1   Stemma of the 13 Dunhuang Chinese manuscripts containing the *Jingangwuliwen* 金剛五禮文 (Source: Yoshida 1994: 362)

8.3.2   Identifying linguistic splits based on texts

The language name "Sogdian" means "language(s) spoken by people in Sogdiana." Despite the impression that the term may give, it should be noted that there was no such thing as "Standardized Sogdian." This is because there never was a unified state covering the whole of Sogdiana (currently Uzbekistan and Tajikistan) prior to Islamization. The Sogdian people were international traders along the Silk Road, and were actively engaged in commercial activities in China and other foreign lands (Map 8-3, Plate 8-3).

Written texts in the Sogdian language differ from one another reflecting such a linguistic situation. In particular, the language of Christian Sogdian texts, written in Syriac script (Plates 8-4 and 8-5), varies considerably from text to text.

Here a question arises as to whether this heterogeneity and diversity reflect different developmental stages, or whether they are regional dialects, varieties that existed at the same time. It appears that both situations are found. There are phenomena which could be interpreted as either, such as difference in the number of case categories and the weakening and/or loss of articles. They could reflect either different developmental stages or the difference between innovative and conservative varieties from the same era. However, as once demonstrated (Yoshida 1980), there are varieties that clearly reflect regional differences. Past progressive forms of the verb are one such example, based on which Christian Sogdian manuscripts can be readily classified into two groups. Details are explained below.

Two past progressive forms are shown in (1) from manuscript C5 (which is a translation of the *New Testament*), and in (2) from manuscript C2 (which is a collection of hagiographies and other miscellaneous texts) respectively. The forms *wāββēk mātāṃt* in the former and *wāββāzaṃt* in the latter both mean "they were speaking". In (1), two verbs, namely *wāββēk* (the present participle of the verb *wāβ* "to speak") and *mātaṃt* (the past
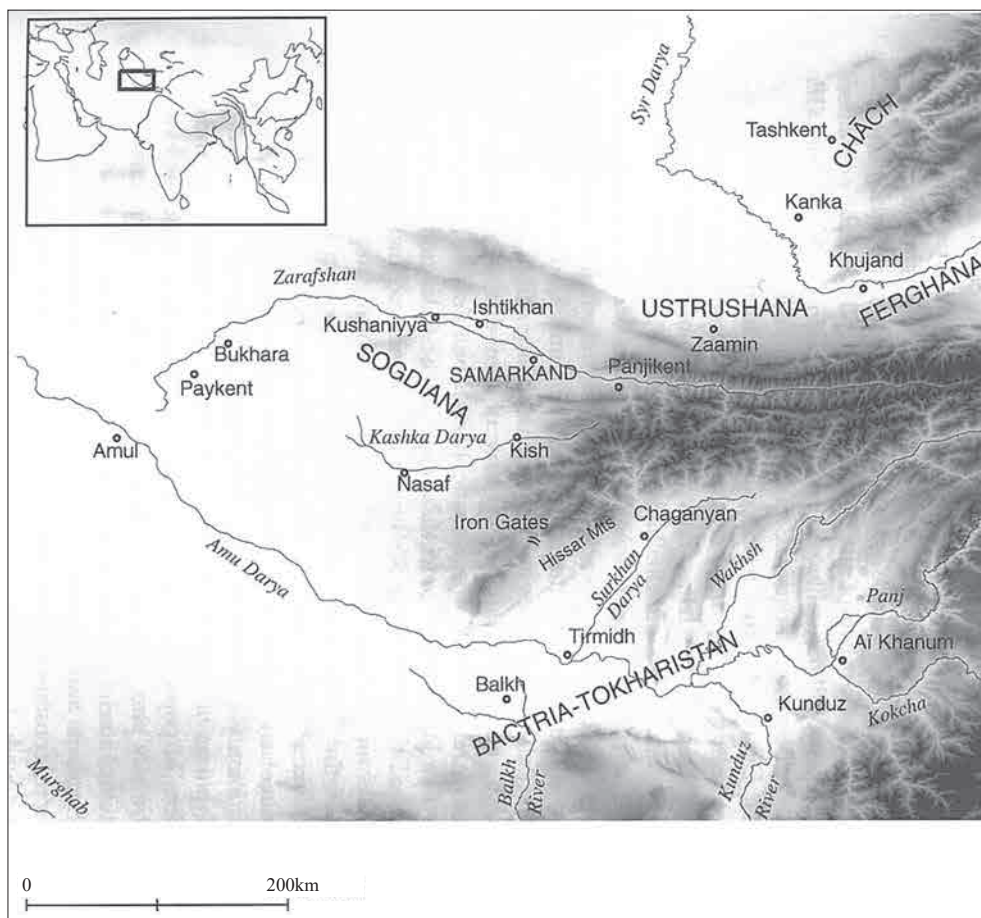
Map 8-3   The location of oasis cities in Sogdiana (de la Vaissière 2005: 15)

form of the substantive verb) are combined to form a periphrastic expression. In (2), a single verb *wāβāzaṃt* is used which consists of *wāβ,* the suffix *āz* denoting the continuous past, and the 3ʳᵈ person plural ending *aṃt.* The original forms in Syrian corresponding to these expressions are identical, a progressive past form of the verb: *'mryn hww* 'they were speaking.'

(1)   *'t*   *w'nw*   *w'byq*   *m'tnt*   *mrtxmyt*
      ət   wānō   wāβēk   mātāṃt   martaxmēt
      and   thus   saying   were.3PL   men
      'People were saying like that.' (Manuscript C5, folio 72 verso 16)

(2)   *c'nw*   *myd*   *w'b'znt*
      čānō   mēδ   wāβāzaṃt
      when   thus   were.saying.3PL

Plate 8-3   Chinese statuette representing a Sogdian caravaneer (8th century) (Qianling Museum 2008: 111)

'When they were saying like that' (Manuscript C2, folio 60 recto 31)

Both types of expression appear frequently, however, never simultaneously in a single text. Therefore, it is reasonable to assume that they developed independently from each other to eventually carry the same function, rather than showing different stages of a single process of development. There are many other differences found between the two manuscripts, most of which are readily explained as reflecting more conservative or innovative stages than the other. However, the differences found in past progressive expressions have to be interpreted as dialectal and not chronological.

Figure 8-2 is a stemma showing the relationship among the varieties of Christian Sogdian as attested in the four groups of manuscripts published before 1980. BST II represents the language of manuscript C2, while NT stands for that of C5. ML I-III and BST I share the periphrastic durative past with C5 and constitute one group against C2 (= BST II). The languages of C5 and ML I=III differ from that of BST I in that the present participle ends in *-ēsk* in BST I against *-ēk* in C5 and ML I-III.

What is interesting about these two past progressive passive expressions is that the earlier embryonic form of each is found in those manuscripts that are older than these

Plate 8-5 Christian Sogdian manuscript, C2: n 32, recto (Depositum der Berlin-Brandenburgischen Akademie der Wissenschaften in der Staatsbibliothek zu Berlin — Preussischer Kulturbesitz Orientabteilung)

Plate 8-4 Christian Sogdian manuscript, C5: n 153, recto (Depositum der Berlin-Brandenburgischen Akademie der Wissenschaften in der Staatsbibliothek zu Berlin — Preussischer Kulturbesitz Orientabteilung)
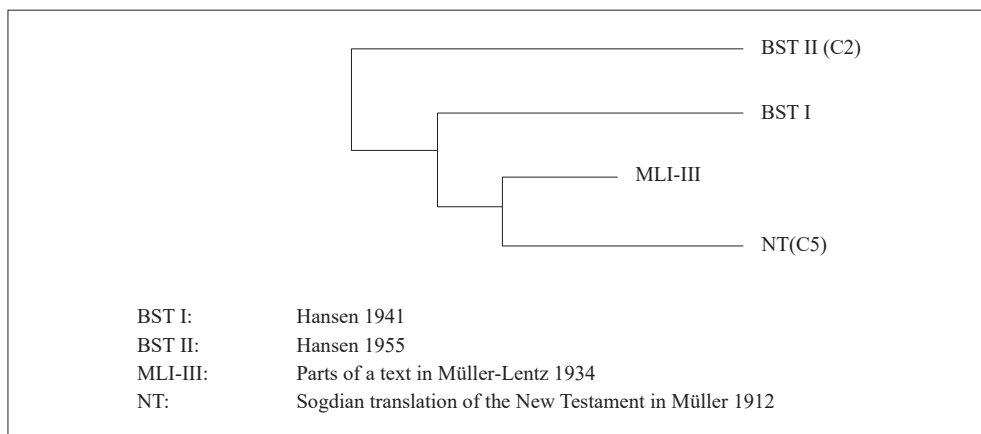
Figure 8-2   Stemma showing the relationship among six Christian Sogdian texts (based on Yoshida 1980: 91)

Christian Sogdian texts. For example, forms such as *ptγwδyy wmᵓtym* [patγōδē wəmātēm] 'we were covering (s.t.)' with the past form of the substantive verb corresponding to (1), and *ᵓskwᵓz* [əskw-āz] 'he stayed' with the progressive past form with a formative *āz* corresponding to (2) appearing in manuscripts older than the Christian Sogdian texts can be identified as the predecessors of the two past progressive expressions discussed above. The occurrence and the usage of these expressions were limited, for there were other forms for the past progressive that were widely used in these older texts. This means that in each of the two Christian Sogdian languages or dialects, one of the several earlier non-productive expressions was inherited to become productive. I consider that this is very convincing evidence to support the idea that there were at least two separate regional varieties in the Sogdian language spoken by Christian Sogdians. In my mind, this is a very good example of capturing a micro-level split of languages through philological research.

The above observations can be summarized as follows. In the proto-stage of a language, there may be alternating forms or expressions that are functionally equivalent. Subsequently, one of them may be selected to be generalized and become productive while others disappear or remain as relics. This selection may take place for multiple features of a language and in different regional or social groups independently from one another. It is the sum of such selections that brings about language split, either into varieties or dialects, and ultimately to become different languages. To understand the mechanism of language change and the process of divergence, it is important to keep in mind that the parent language may have had different linguistic characteristics, any one of which could have subsequently developed to become the major feature in a daughter language. It is not possible to provide a general explanation as to why languages develop new forms from earlier functional equivalents. In order to know that, each case has to be examined in its own context considering the factors that must have influenced the change.

Stemmas that philologists deal with include also those of scripts and numeral signs.

However, space does not allow me to talk about them here.

8.4. Family Tree Models of Indo-European Languages

In this section, different types of models that are used to reflect the relationship among Indo-European languages are summarized and problems associated with them are discussed.

The very first tree model presenting the relationship of Indo-European languages was invented by August Schleicher, and has been reprinted in many text books (Figure 8-3). Problems associated with his tree have been pointed out since that time. First, Schleicher considered languages to be organisms just like biological creatures and thus regarded them as entities that evolved independently of speakers. No language, however, exists without its speakers and therefore there is an obvious problem in this assumption. Second, his model, proposed before the era of the Neogrammarians, reflects the mid-nineteenth century unsophisticated view of Indo-European languages. Third, the significant fact that languages, even after they have split, influence each other through contact could not be adequately reflected. The third point triggered the proposal of an alternative model, namely the Wave model by Johannes Schmidt (1872). Regarding Indo-European languages, a diagram showing linguistic phenomena observed across different branches was subsequently proposed by Schrader (1907), see Figure 8-4. His diagram became commonly known among linguists after Leonard Bloomfield (1933) cited it, see (Figure 8-5). A revised version was published in Hock (1999) see (Figure 8-6).

Linguists have continued to use the family tree model to show relationships among the Indo-European languages. A family tree with about ten branches directly diverging from Proto-Indo-European is the commonly used diagram today (Figure 8-7), in which the interrelationship among the branches are not considered; most of the Indo-European family trees so far proposed are drawn reflecting the geographical distribution of the languages. Exceptions to this are the Indo-Iranian and Balto-Slavic languages.

Since the turn of the twentieth-first century, computational analyses of languages have become more common, where methodologies developed in the study of genetics are applied. Being somewhat reminiscent of Schleicher's tree model, researchers in such areas are seriously tackling the question as to how languages genetically evolve on a large scale. Gray and Atkinson (2003, Figure 8-8) and Nakhleh *et al.* (2005, Figures 8-9a, b) are examples. The latter gives consideration to changes that were incurred by contact. It should be pointed out, however, such computational analyses are still dependent on manual selection by historical linguists depending on their basic knowledge, such as sound changes, existence or non-existence of certain conjugational categories, selection of lexical items, and their assessment of the relative importance of these for the establishment of the relative chronology of branching. It should also be pointed out that when considering contact relationship among groups of genetically related languages in prehistory, the evidence is not exclusively linguistic but outside evidence is also taken into consideration. In particular, information acquired through archaeological research and excavated materials are referred to regarding peoples' movements.
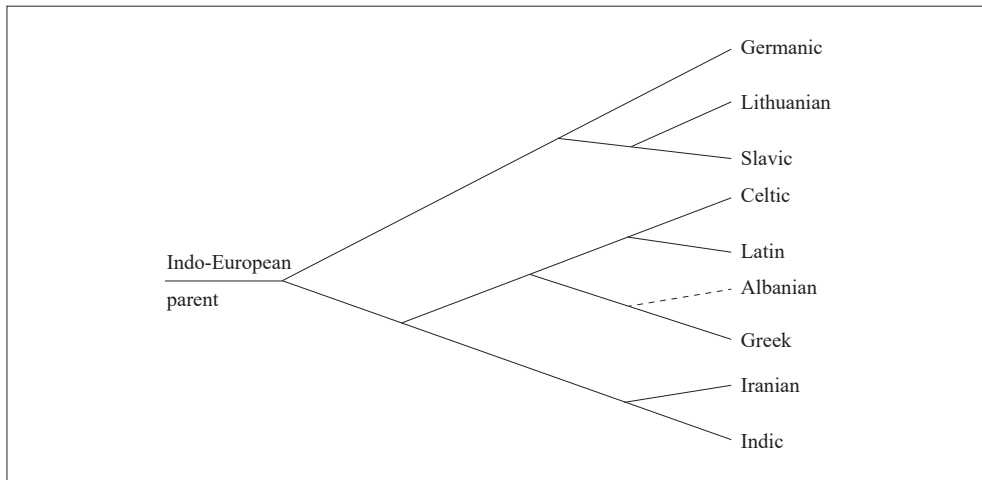
Figure 8-3   Schleicher's family tree model of the Indo-European languages (Schleicher 1861) after Clackson (2007: 11)
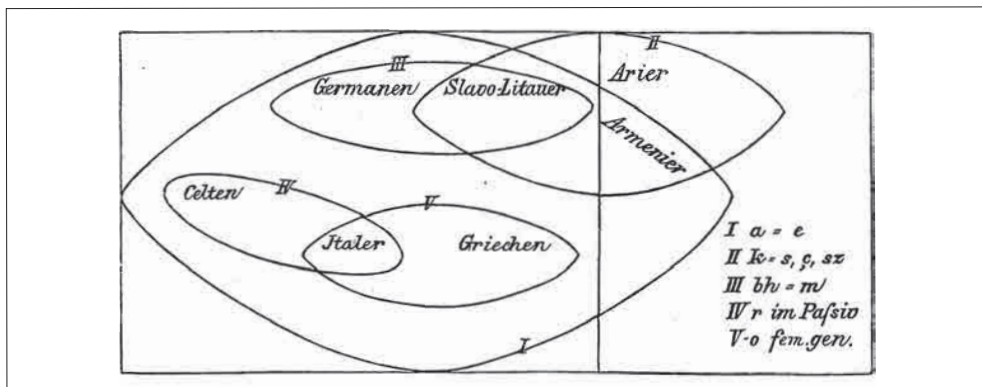

Figure 8-4   Schrader's wave model (Schrader 1907: 65)

Genetic relationship among Indo-Iranian languages is considered to be relatively straightforward, though not without problems. There is a language called Nuristani, for example, whose affiliation is not clear. This language is spoken in the mountainous area of Afghanistan, and it is clearly an Indo-Iranian language. However, which sub-branch it belongs to has not been identified. The Proto-Indo-Iranian language is known to have split into two, namely Iranian and Indo-Aryan. Nuristani could belong to either of these, or otherwise could form a third branch. In the most recent examination of the position of Nuristani, Degener (2003: 112) states that the position of this language could not be identified solely based on linguistic comparison. Taking the results of archaeological research, she concludes that Nuristani belongs to the Indo-Aryan sub-branch, and that it split off at an early stage and was continuously exposed to Iranian languages of the other sub-branch. Degener's claim is based on the locations of archaeological sites in which
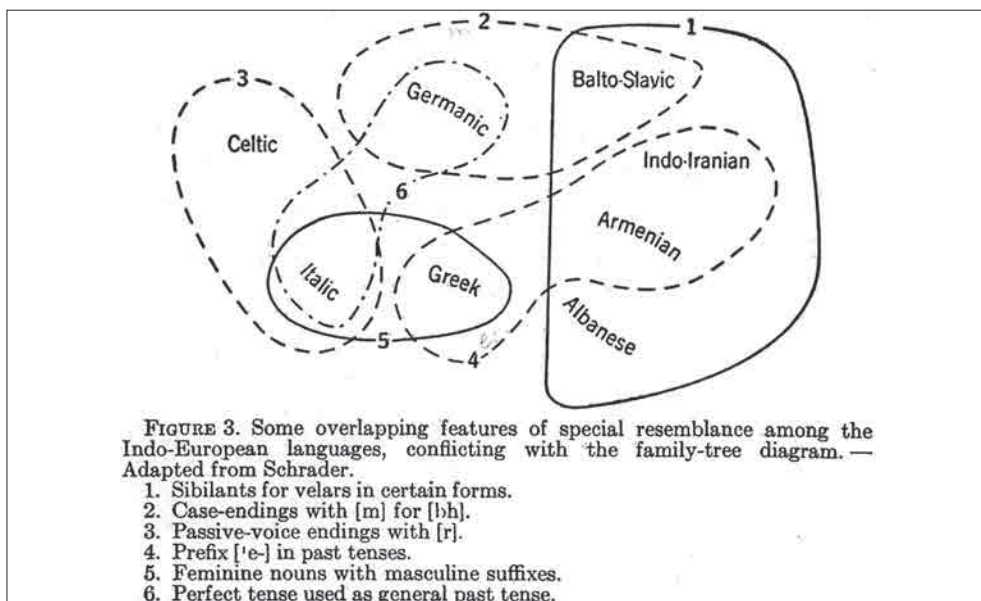
FIGURE 3. Some overlapping features of special resemblance among the Indo-European languages, conflicting with the family-tree diagram. — Adapted from Schrader.
1. Sibilants for velars in certain forms.
2. Case-endings with [m] for [bh].
3. Passive-voice endings with [r].
4. Prefix ['e-] in past tenses.
5. Feminine nouns with masculine suffixes.
6. Perfect tense used as general past tense.

Figure 8-5   Wave model in Bloomfield 1933 based on Schrader (Bloomfield 1933: 316)
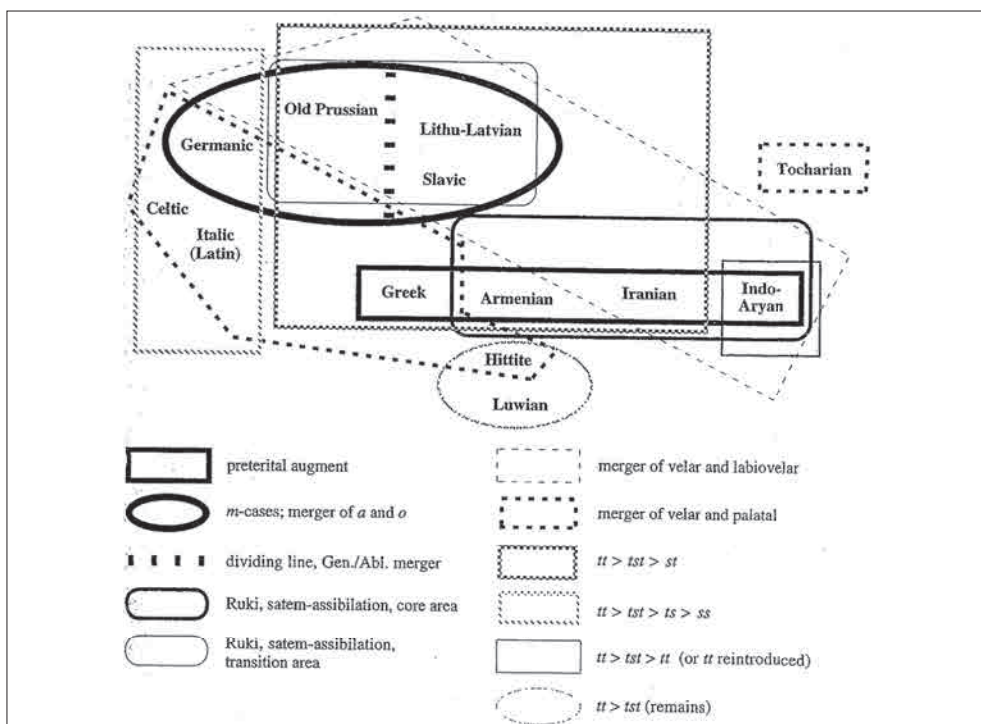

Figure 8-6   Hock's revised wave model of the Indo-European languages (Hock 1999: 15)
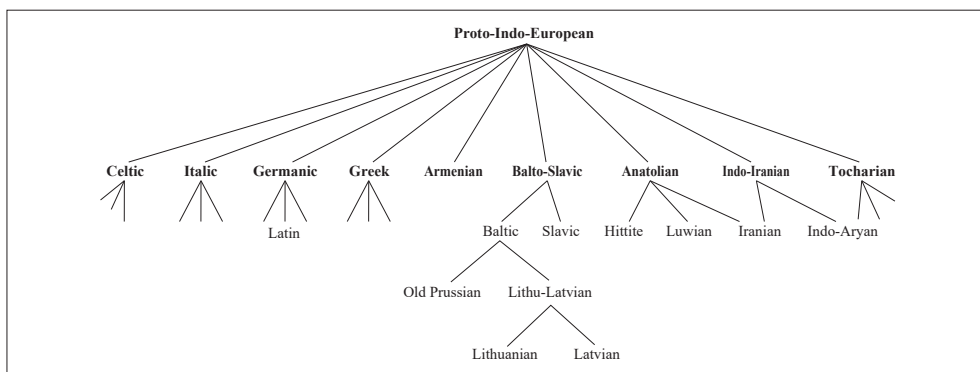
Figure 8-7   Hock's family tree model of the Indo-European languages (Hock 1999: 14)

the absence of graves specifically points to Proto-Iranian culture. Her tree diagrams showing the position of Nuristani (Figure 8-10) has a unique form, where influence through language contact is expressed by bent lines.

8.5. Sogdian and the Family Tree Model

In this section, issues related to the family tree model of Iranian languages, specifically those related to Sogdian, will be discussed.

Various family trees have been proposed hypothesizing the relationship among Iranian languages. Relatively recent ones are shown in Figure 8-11 and Figure 8-12, neither of which has been drawn by a specialist of this language group. Researchers of Iranian languages are not particularly keen in investigating the details of the genetic relationship among the languages, and it appears that they are content with the rough classification presented in these diagrams.

As can be seen in Figure 8-11, it is generally accepted that the Iranian languages are classified into two main groups, West Iranian and East Iranian, a subgrouping hypothesis supported by phonological innovations. (Figure 8-12 is unique and radically different from the generally accepted traditional view; the evidence for it as well as the basis for the exact dating of the time of branching elude the present author.) West Iranian is further classified into the North-West Iranian and South-West Iranian, again the subgrouping is supported by phonological innovations. East Iranian is sometimes classified into North and South also, however, this is based on the geographical distribution of languages rather than on linguistic innovations. Some linguistic examination concerning the relationship among these languages is presented below.

8.5.1 Sogdian and Yaghnobi

The focus of this subsection is the relationship between the Sogdian language documented in old manuscripts and Yaghnobi or Modern Sogdian. Figure 8-13 is an attempt by the present author to show the relationship among different stages of Iranian
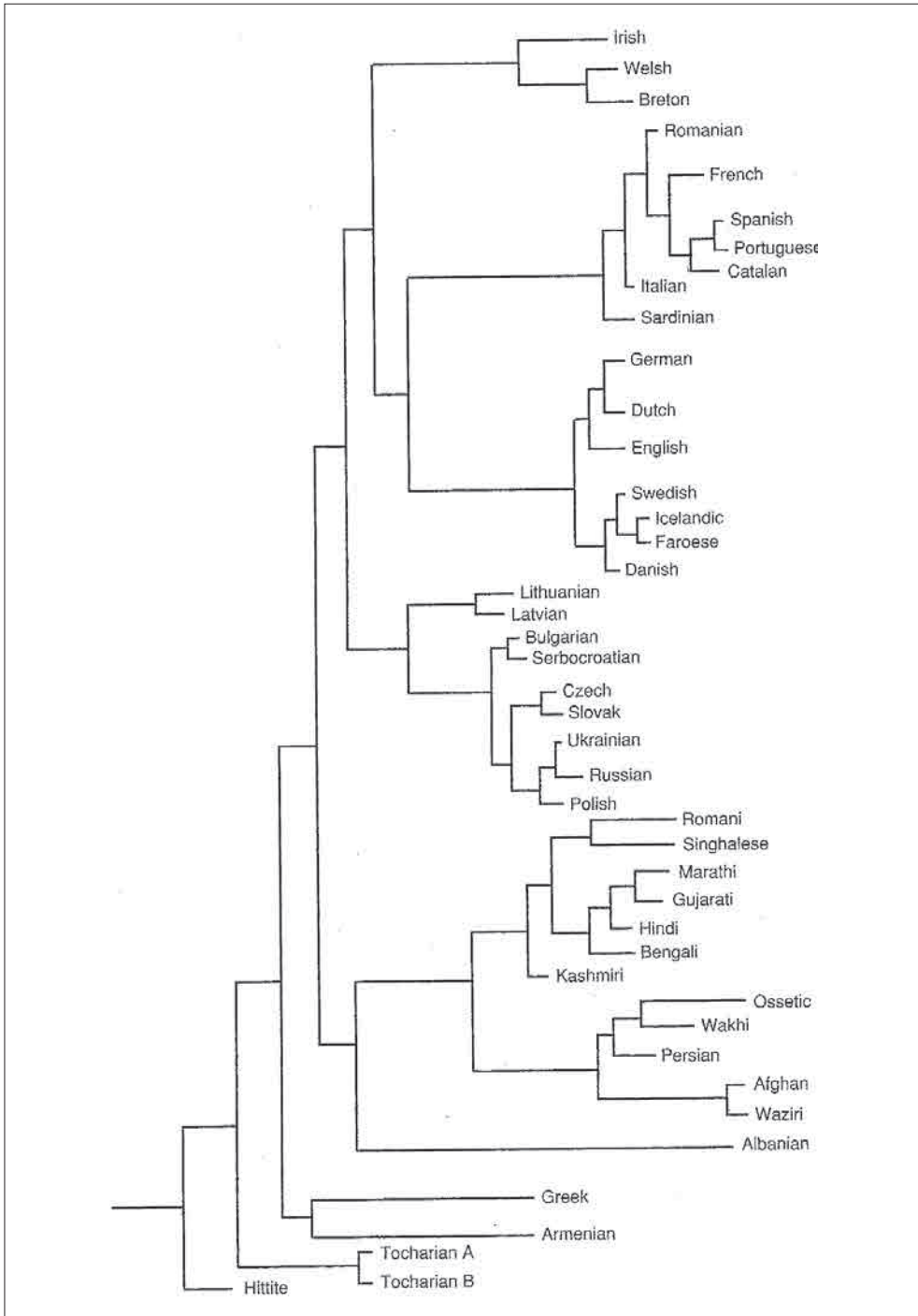
Figure 8-8   A family tree model of the Indo-European languages proposed by Gray and Atkinson (2003: 435–439) after Clackson (2007: 11)
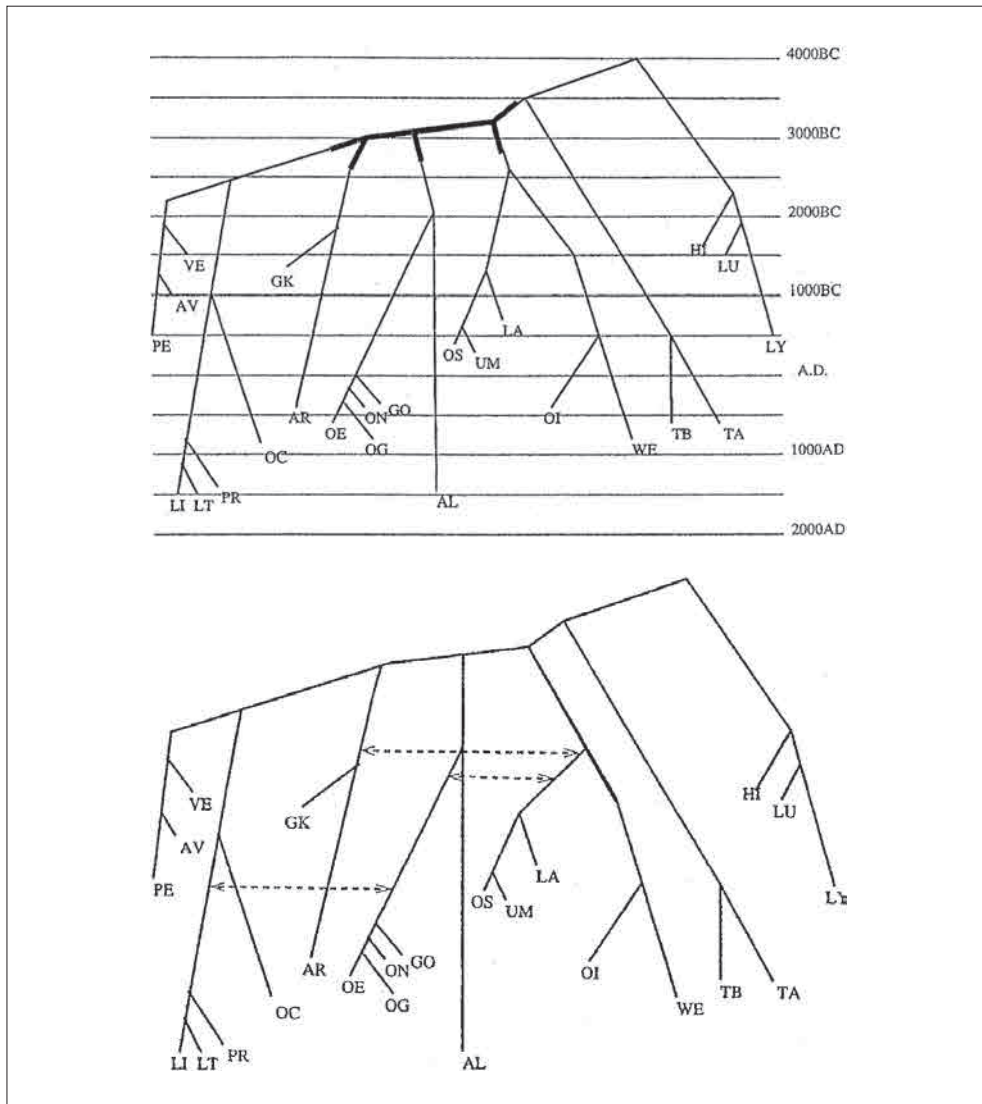
Figure 8-9a, b   A family tree model of Indo-European languages proposed by Nakhleh *et al*. (2005)

languages (Ancient, Middle, and Modern), including Sogdian and Yaghnobi. If we assume that Proto-Indo-Iranian broke into two or three branches about 3,500 years ago, the split into East Iranian and West Iranian would have occurred about 3,000 years ago, when the ancestors of Modern Persian and Yaghnobi seem to have diverged.

Yaghnobi and Sogdian "resemble" each other and the former is sometimes referred to as Modern Sogdian by those who study Iranian. To confirm the fact that the two languages are linguistically closer to each other than to Modern Persian, the basic vocabularies of Modern Persian, Sogdian and Yaghnobi are compared. The comparison is
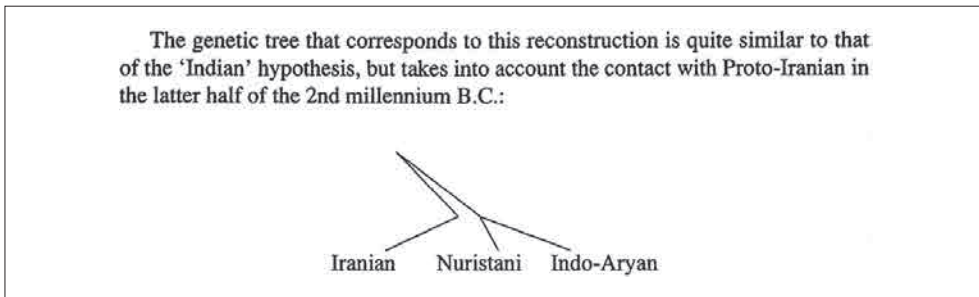
The genetic tree that corresponds to this reconstruction is quite similar to that of the 'Indian' hypothesis, but takes into account the contact with Proto-Iranian in the latter half of the 2nd millennium B.C.:

Iranian   Nuristani   Indo-Aryan

Figure 8-10   The position of Nuristani according to Degener (2003: 116)

expected to show that Yaghnobi shares a larger number of similar lexical items with Sogdian than it does with Persian. First, Swadesh's 100 wordlist was used. As expected, in some sets, all forms are cognate among the three languages (e.g., 3 in Table 8-1), in some other sets, the forms in the three languages have different sources (e.g., 13 in Table 8-1), and in the others, cognates are found in either two of the three languages (e.g., 20, 21, 28 in Table 8-1). The result of the comparison is that 64 apparent cognate sets are found between Sogdian and Yaghnobi, and 61 apparent cognate sets are found between Sogdian and Modern Persian. As mentioned earlier, Sogdian appears to be much more similar to Yaghnobi than it is to Modern Persian and this result is contrary to what one would expect. One would expect many more cognates between Sogdian and Yaghnobi than between Sogdian and Modern Persian.

The same examination was conducted using the Leipzig-Jakarta List, hereafter LJ List (Tadmor et al. 2010). The LJ List contains 100 words that are considered to be most resistant to borrowing. While the items in the Swadesh list were intuitively selected, the LJ List was compiled based on a statistical analysis of 1,500 words in 41 languages from across the world. Considering the fact that Yaghnobi has undergone intensive borrowing from Tajik (a variety of Modern Persian) and Uzbek spoken in surrounding areas, the LJ List appears to be more reliable for our purposes. The two lists of basic vocabulary differ from each other, and 38 words are not shared by the two sets. When one compares the 38 Sogdian words belonging to the Swadesh list with those of Yaghnobi, one finds 27 apparent cognates among them. In the case of the 38 items of the LJ list, only 19 (or 21 depending on how one counts) words turn out to be cognates. This result is astonishing again, because one would expect a "better performance" from the LJ list, based on modern technology and advanced linguistic knowledge.

This raises a serious question as to the liability of the JL List. Tadmor *et al*. (2010) classify 1,500 words in the 41 languages into the following five categories: i) Clearly borrowed; ii) Probably borrowed; iii) Perhaps borrowed; iv) Very little evidence for borrowing; v) No evidence for borrowing. However, the criteria as to how each form is classified into one of these may have to be reevaluated. For example, Japanese *sekai* (世界) 'the world' is classified as i), while its synonym *yo* "(this) world" is classified as v). These are reasonable, for it is known that the former is known to have been borrowed from Chinese, while the latter is a word that originated in Japan. However, it seems
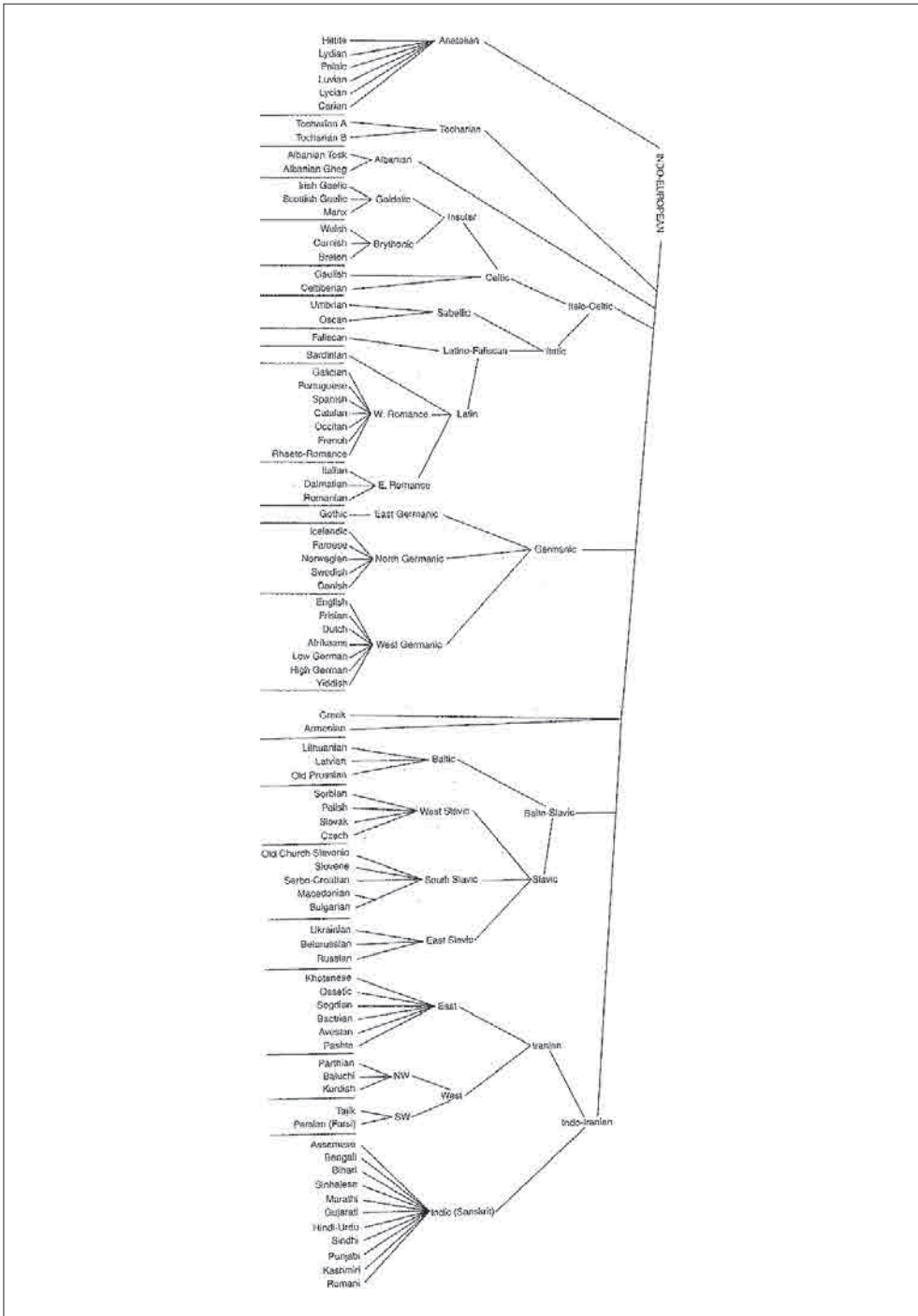
Figure 8-11   A family tree showing the genetic relationship of Iranian languages (I) (Campbell and Poser 2008: 84–85)
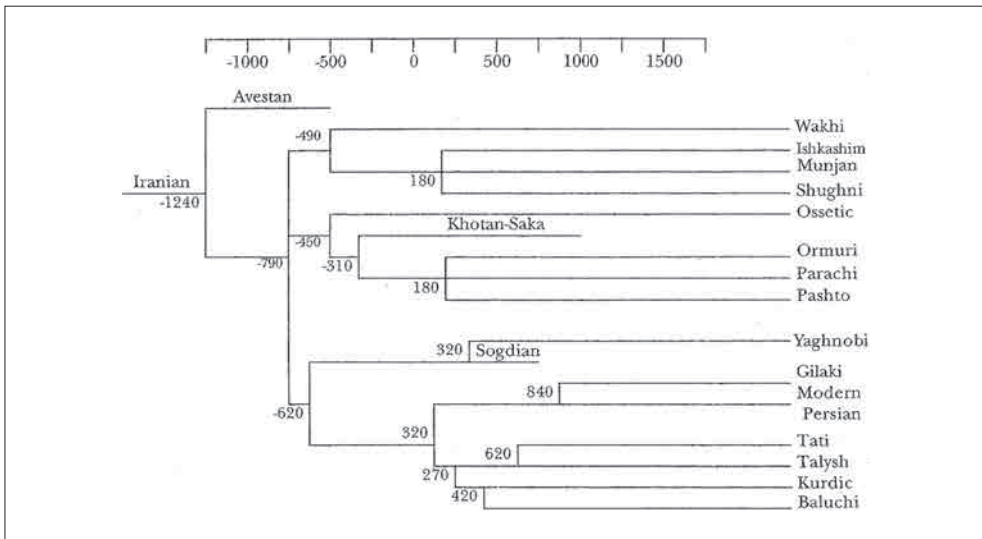
Figure 8-12   A family tree showing the genetic relationship of Iranian languages (II) (Blažek 2007)
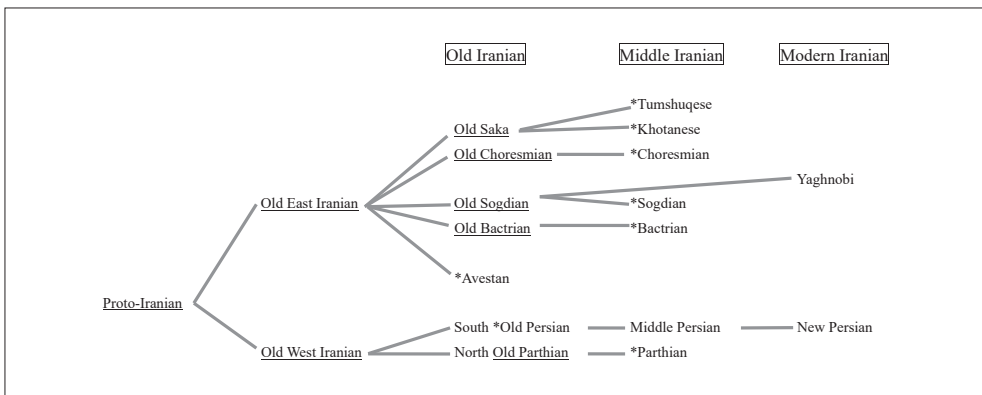
Figure 8-13   The family tree of Ancient, Middle, and Modern Iranian languages including Yaghnobi and New
Persian (compiled by the author)
Note: The names of unattested languages are underlined, e.g., Proto-Iranian; languages attested in
written texts are indicated by an asterisk (*), e.g., *Old Persian.

Table 8-1   Comparison of some words in Sogdian, Yaghnobi and Persian[1] (prepared by the author)

Gloss	Sogdian	Yaghnobi	Persian
3. we	māx	mox	mā
13. big	mazēx	katta	buzurg
20. bird	mərɣ-	paranda	murɣ
21. dog	əkut-	kut	sag
28. skin	čarm	pust	pūst

preposterous that *gādoman* (ガードマン, 'a security guard'), a Japanese word formed based on two English loanwords, *gādo* 'guard' and *man* 'man', is classified as iv), *nyūyoku* (入浴, 'bathing, to have a bath') as iii), and *shio* (塩, 'salt') and *suijōki* (水蒸気, 'steam, vapor') as ii); while *shio* is a good original Japanese word, *nyūyoku* and *suijōki* are no doubt Chinese borrowings as their everyday spelling in Chinese characters and their pronunciations clearly show. As far as one can see from the classification of these examples of Japanese lexical items, the LJ list is far from reliable.

In any case the above statistics do not explain the similarities that are intuitively "felt" between Sogdian and Yaghnobi, and tell us that to identify where this intuitive feeling comes from, more than such simple statistics as checking the number of shared vocabulary is required. In the case of Sogdian and Yaghnobi, a detailed examination shows that the source of the "similar" impression between the two languages is the shared systems rather than the number of shared lexical items. For example, morphological alternations involving the verb 'to go' is clearly shared between Sogdian and Yaghnobi, but not with Persian. The forms for 'to go' in the three languages, when the finite forms are compared, are all cognates; in fact, they all show the same form, namely *šaw-*. However, when we look at the pattern of their derivational systems, it turns out that the same alternation is shared between Sogdian and Yaghnobi but not by Persian. The forms in Sogdian and Yaghnobi are *šaw-* 'to go (present)' and *īt-* 'to go (past)' and *šaw-* 'to go (present)' and *eta-* 'to go (past)' respectively. However, in Persian, the present stem is *šaw-* and the past stem *šud-*. This means that, to adequately capture shared characteristics of two (or more) languages, simply counting the number of cognates is not enough, and shared anomalies of the morphological and other systems need to be examined. In the example given above, whether the suppletion is shared or not appears to draw a clear line among the three languages, one grouping Sogdian and Yaghnobi against Persian. And, when two related languages share a bundle of irregular systems such as suppletion, probably this gives the impression that the languages are more "similar" than other genetically related languages.

8.5.2 "Grimm's Law" of the East Iranian languages

In this subsection, two further questions are discussed that are related to the position of Sogdian in the family tree. They both have to do with the qualification of certain shared innovations as a criterion for subgrouping.

The first topic is the change of earlier voiced stops to fricatives. In all the Iranian languages, voiced stops became fricatives after a vowel. In addition, in East Iranian languages, they were spirantized before a vowel as well. Examples of sound correspondences of **b-*, **d-*, and **g-* are shown in Table 8-2, where the corresponding fricatives are presented by Greek letters β, δ, γ following the Iranian linguistics convention. Similarly, Old Iranian **-ft-* and **-xt-* became **-βt-* and **-γt-* respectively in East Iranian. The fact that the sounds in question were pronounced as fricatives rather than stops is clear from the spellings in Manichaean and Arabic scripts, which distinguish voiced fricatives from the corresponding stops by diacritics.

The phonemic change or phonemic shift shown above, which could be referred to as

Table 8-2 Comparison of the western and eastern Iranian reflexes of Old Iranian *b-, *d-, *-ft-, and *-xt- (prepared by the author)

Gloss	Old Iranian	Avestan	West Iranian	East Iranian			
			New Persian	Bactrian	Sogdian	Yaghnobi	Choresmian
'bear'	*bara-	bara-	bar-	cf. βurd	βar-	var-	βar-
'hold'	*dāra(ya)-	dāra(ya)'	dār-	lēr-	δār-	dor	δāray
'cow'	*gāu-	gāu-	gāw	γāw	γāw	γow	γāw
'seven'	*hafta	hafta	haft	---	aβta	avd/aft	aβd
'given'	*baxta-	baxta-	baxt	βaγd	βaγt	---	βaγd

'East Iranian Grimm's law,' is shared by all Eastern Iranian languages, and it appears as though the Iranian languages could be classified into East and West with little problem. Nevertheless, it is interesting to note that this sound change is not attested in the language of the *Avesta*. Although we still do not know where in Iran Avestan was spoken, the language is generally assumed to belong to the East Iranian group, because all the place names mentioned in the texts belong to East Iran, and because some of its linguistic features are shared by other East Iranian languages. This tells us that the change took place in a relatively late stage of the development of the East Iranian languages. Thus, the phenomenon is to be analyzed as drift rather than shared inheritance. It would not be possible to assume that this type of systematic change took place due to language contact. The distribution of the fricative consonants can be referred to as forming one of the isoglosses indicating the geographical boundary between East and West Iranian. Nevertheless, it should be regarded as a change which followed but not preceded the divergence of East and West Iranian. If we had no Avestan texts, we would have had no hesitation drawing a family tree of the Iranian languages where the parent language splits into two branches, East and West, by this "East Iranian Grimm's Law". The observation presented here makes us even question if the phenomena understood as Grimm's Law and Verner's Law could in fact be drift. The point is that what a family tree model presents does not always exactly match what actually happened in languages in the past.

Since Schmidt's proposal of the wave theory, linguists have been emphasizing that there exist languages to which the family tree model may not be easily applied. For example, Bloomfield states, referring to linguists who employ a family tree model: "They accepted the uniformity of parent languages and their sudden and clear-cut splitting, as historical realities." (Bloomfield 1933: 311)

This perception has been shared by more recent linguists as well. For example, Nakhleh *et al.* (2005) note:

> But this approach obviously cannot be relied on to reconstruct evolutionary histories for those language families in which related dialects have evolved in close contact with each other. (Nakhleh et al. 2005: 388)

It is known that there was no sudden and clear-cut split that took place in Iranian languages, and this is true with other languages of the world. Austronesian languages spoken in an island environment may have been, but even so not always. Even when speakers migrated from an island to another island, they still could have gone back to the original islands and kept contact with the groups of speakers there, although the present author is ill-qualified to comment on this matter. It should be noted in passing that linguistic change in the actual world never takes place in a lab-like environment. I opine that researchers should be always questioning the appropriateness of using a model that presumes a lab-like situation.

From a linguistic philological point of view, where the history of individual languages documented in written records are examined, the realities of the changes that each member language undergoes are all too unique and specific to meet the generalization that is required for applying a family tree model. This echoes the words of Jules Gilliéron who established a methodology of classical dialect geography against that of comparative linguistics, "*Chaque mot a son histoire* (Each word has its own history)." In our context, this could be restated as "*Chaque langue a sa histoire* (Each language has its own history)." Having said that, it is not my intention to completely deny the use of the family tree model. The model, including stemmas for classifying manuscripts mentioned earlier in this paper, is useful in some research contexts when applied with precaution.

8.5.3 Shared innovation and shared inheritance

Most textbooks of historical linguistics tell us how to draw a phylogenetic diagram of languages. The nodes are determined based on shared innovations, and not on shared inheritance. For example, "It is now generally agreed among linguists that the most certain sub-groups are reconstructed on the basis of unique shared morphological innovations." (Clackson 2007: 5–6)

I would like to point out that there are some known cases where this principle does not hold. Here, we will focus on Sogdian, Yaghnobi, and Choresmian, which all belong to the East Iranian branch (Figure 8-13), Choresmian is spoken in areas adjacent to Sogdian. As background information, East Iranian languages are not likely to have formed dialect continuums, for most of them were spoken in areas where little habitable lands, such as in high mountains and deserts, interrupt everyday communication between language communities. Language contact nevertheless must have occurred in such occasions as invasion, or along the routes of caravans and/or nomads, and when people migrated as a result of population increase. However, details of such contact are not known well enough.

According to the subgrouping hypothesis shown in Figure 8-13, Sogdian and Yaghnobi belong to the same sub-branch while Choresmian to another. However, Sogdian and Choresmian share some characteristics in common, which can be identified as Clackson's shared morphological innovations, as opposed to Yaghnobi. For example, like Romance and Germanic languages, in both languages a past perfect tense conjugation developed where the auxiliary verb *δār*- 'to have, hold' forms part of it. Examples are

shown in (3) and (4), where the second element of each sentence is the auxiliary verb. This system is found only in these two languages among all the Iranian languages and it is not shared by Yaghnobi either.

(3) Sogdian

    'krtw   *δ'r'm*
    əktu-   δārām
    'I have done.'

(4) Choresmian

    ktk   *δ'ry'm*
    əktək   δārayāmi
    'I have done.'

Another feature shared by Sogdian and Choresmian is the formation of imperfect past forms. It is known that in Old Iranian, the imperfect past verb stem was formed by a derivation whereby the vowel *a-* is added to the beginning of the present tense stem of the verb. Among the daughter languages, this system is retained almost exclusively in Choresmian, Sogdian, and Yaghnobi. However, the formation process differs depending on the language. In Yaghnobi, like in Old Iranian, the imperfect past is formed by adding the vowel *a-* to any verb. However, in Sogdian and Choresmian, *ā-* is inserted after the initial consonant of verbs starting with a consonant cluster, whereas the consonant *m-* is added to those verbs starting with a vowel. (Here the rules are presented in a somewhat simplified form.)

(5) Derivation of the imperfect past stem in the three Iranian languages

Sogdian	θβar-	'to give'	> θāβar-
	āγāz-	'to begin'	> m-āγāz-
	anxwāy	'to break'	> m-anxwāy-
Choresmian	hβar-	'to give'	> hāβar-
	āγāz-	'to begin'	> m-āγāz-
Yaghnobi	tifar-	'to give'	> a-tifar-
	unxōy-	'to break'	> a-unxōy-

N.B. In Choresmian the verb corresponding to Sogdian *anxwāy* has not been attested, while the Yaghnobi counterpart of Sogdian *āγāz-* is not known.

On the other hand, there are features shared between Choresmian and Yaghnobi but not by Sogdian. One of the Old Iranian third person plural endings is known to have been *-r* and this is retained in the former two languages, but not in Sogdian, see Table 8-3.

One may wonder if the lack of the *-r* suffix in Sogdian is accidental, for Sogdian is an extinct language. However, this is clearly not the case. First, there are a good number of manuscripts of the language still available. Second, third person plural is the person

Table 8-3   Third person plural suffixes in three Iranian languages (prepared by the author)

	'they did'	'they held'
Sogdian	kun-aṃt	δār-aṃt
Choresmian	m-ak-āra	δāray-āra
Yaghnobi	a-kun-or	a-dor-or

and number that occurs rather frequently in the literature. Therefore, it is unlikely that the *-r* suffix had in fact existed in the language without appearing in any of the manuscripts available today.

The *-r* suffix is known to be an archaic feature, and the *-r* ending existing in Choresmian and Yaghnobi appears to be a typical example of shared inheritance, whereas Sogdian had innovated in this respect. However, the occurrence of the *-r* suffix in Yaghnobi is limited to the past tense, while in Choresmian, it occurs in all tenses and moods. Looking at other languages in East Iranian, in Khotanese, which retains the old distinction of active and middle voices, the distribution of the *-r* ending seems to have been generalized and occurs in all middle forms except for the indicative mood. Bactrian has no *-r* endings like Sogdian. Should this situation be analyzed as Sogdian and Bactrian sharing the innovation of the loss of the *-r* ending, or the other three languages share the characteristic of inheriting the *-r* with the modifications not shared among them?

If we follow the instruction in textbooks, the distribution of the *-r* suffix would not be considered as a feature that could be used for identifying the genetic relationship of languages, because it is a shared inheritance rather than a shared innovation. However, considering such cases as the ones presented in this subsection, I wonder if that is a correct approach. To conclude, what appears to be a simple shared inheritance deserves careful attention as criteria for language classification. As for the isoglosses that are not matching among Choresmian, Sogdian, and Yaghnobi, they remain unexplained. This was to show that it should be always kept in mind that cases described in textbooks are clear, easy to comprehend, and occasionally normalized (or simplified) ones, but the realities are far more complicated.

8.6. Conclusion

The tree diagram is a useful tool for linguistic analysis and is widely applied to show genetic or typological relationships of various groups of languages. However, continuously dealing with rather trivial but real linguistic changes documented in the written records of ancient linguistic communities from day to day makes the present author aware that the facts are not well captured by tree diagrams. In such a situation, one would be forcing the data, if one tries to draw a tree reflecting the details, possibly selecting data to meet the model. However hypothetical, a phylogenetic diagram, sometimes presented even with absolute chronologies by brave and ambitious linguists, would be easier for lay people to comprehend than the real but boring facts about languages. A false impression can be conveyed to the public that the tree represents

scientific facts which have been agreed upon by specialists. It should be kept in mind that, when a linguist tries to attach more historical reality to a family tree than can be scientifically shown based on language data, such work would not only be useless but also detrimental to those who are ignorant of linguistics.

Note

1) The Greek letter γ is used to indicate [ɣ] following Iranian linguistic conventions.

References

Bailey, H. W.

    1971   Ancient Kamboja. In C. E. Bosworth (ed.) *Iran and Islam: In Memory of the Late Vladimir Minorsky*, pp. 65–71. Edinburg: Edinburgh University Press.

Beal, S. (trans.)

    1906   *Si Yu Ki: Buddhist Records of the Western World. Translated from the Chinese of Hiuen Tsiang, AD 629*. London: Kegan Paul Trench, Trübner.

Blažek, V.

    2007   From August Schleicher to Sergei Starostin: On the Development of the Tree-diagram Models of the Indo-European Languages. *The Journal of Indo-European Studies* 35(1–2): 82–109.

Bloomfield, L.

    1933   *Language*. New York: Holt.

Campbell, L. and W. J. Poser

    2008   *Language Classification: History and Method*. Cambridge: Cambridge University Press.

Clackson, J.

    2007   *Indo-European Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Collins, B. A. [translation of Al-Muqaddasi]

    1994   *The Best Divisions for Knowledge of the Region. Ahsan al-Taqasim fi Ma'rifat al-Aqalim*. Reading: Garnet Publishing Limited.

Degener, A.

    2003   The Nuristani Languages. In N. Sims-Williams (ed.) *Indo-Iranian Languages and Peoples*, pp. 103–117. Oxford: British Academy.

Dixon, R. M. W.

    1997   *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

Frye, R. N.

    1984   *The History of Ancient Iran*. Munich: Beck.

Gray, R. D. and Q. D. Atkinson

    2003   Language-tree Divergence Times Support the Anatolian theory of Indo-European Origin. *Nature* 426(6965): 435–439.

Hansen, O.

   1941   *Berliner soghdische Texte I (Abhandlungen der Preussischen Akademie der Wissenschaften, Phil.-hist. Kl., No. 10).* Berlin: Akademie Verlag.

   1955   *Berliner soghdische Texte II (Akademie der Wissenschaften und Literatur, No. 15).* Berlin: Akademie Verlag.

Hock, H. H.

   1999   Out of India? The Linguistic Evidence. In J. Bronkhorst and M. M. Deshpande (eds.) *Aryan and Non-Aryan in South Asia: Evidence, Interpretation and Ideology*, pp. 1–18. New Delhi: Manohar Publishers and Distributors.

Müller, F. W. K.

   1912   *Soghdischer Texte* I (*Abhandlungen der Preussischen Akademie der Wissenschaften, Phil.-hist. Kl., No. 2*). Berlin: Akademie Verlag.

Müller, F. W. K and W. Lentz

   1934   *Soghdische Texte* II (*Sitzungsberichte der Preussischen Akademie der Wissenschaften, Phil.-hist. Kl., No. 21*). Berlin: Akademie Verlag.

Nakhleh, L., D. A. Ringe, and T. Warnow

   2005   Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages. *Language* 81(2): 382–420.

Qainling Museum

   2008   *Exotic Flavor of the Foreigners on the Silk Road: Terracotta Hu Man of the Tang Dynasty.* Beijing: Wenwu Publisher.

Schleicher, A.

   1861   *Compendium der vergleichenden Grammatik der indogermanischen Sprachen.* Weimar: Hermann Böhlau.

Schmidt, J.

   1872   *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar: Hermann Böhlau.

Schmitt, R.

   1989   *Compendium Linguarum Iranicarum.* Wiesbaden: Reichert Verlag.

Schrader, O.

   1907   *Zur Geschichte und Methode der linguistisch-historischen Forschung*. Bremen: H. Costenoble.

Sima, Q

   1993   *Records of the Great Historian: Han Dynasty II*, translated by B. Watson. 2nd ed. New York: Columbia University Press.

Tadmor, U., M. Haspelmath, and B. Taylor

   2010   Borrowability and the Notion of Basic Vocabulary. *Diachronica* 27(2): 226–246.

de la Vaissière, É.

   2005   *Sogdian Traders: A History*, translated by J. Ward. Leiden: Brill.

Yang, H.-S., Y.-H. Jan, S. Iida, and L. W. Prenston (eds. and trans.)

   1984   *The Hye Ch'o Diary: Memoir of the Pilgrimage to the Five Regions of India*. Berkeley and Seoul: Asian Humanities Press.

Yoshida, Y.

1980   Kirisutokyou Sogudo-go no Hougen ni Tsuite (On the Dialectology of Christian Sogdian). *Orient* 23(1): 83–93. [In Japanese]

1994   Sogudo Moji de Hyoukisareta Kanjion (Chinese in Sogdian Script). *The Toho Gakuho* (*Journal of Oriental Studies, Kyoto*) 66: 271–380. [In Japanese]