Chapter 6

The classification of the Tungusic languages

Lindsay J. Whaley and Sofia Oskolskaya

Abstract

This chapter surveys previous attempts to classify the genetic relationships among the Tungusic languages. We examine the set of sound correspondences that can be employed in this classification and argue that, if one assumes binary branching for a cladistic classification, there are three plausible classifications that result from the application of the classical comparative method. Next, a Bayesian phylogenetic analysis of basic vocabulary is undertaken to determine whether that analysis provides any further evidence for which of the three classifications is preferred. We conclude that it does and that one of the best classifications of Tungusic places Manchu, Xibe, and Jurchen in a Southern Branch together with Udihe and Nanai complexes and the Even-Evenki complex in a Northern Branch. Though our analysis does not exclude the most common classification in which the Manchuric branch separated first from all other Tungusic languages.

Keywords: Tungusic, comparative method, classification, Bayesian phylogenetic methods

6.1 Introduction

The twelve or so languages that comprise the Tungusic language family are spoken across a vast portion of Central and Eastern Asia. To the north, one finds speakers of Even and Evenki up in the Arctic Circle, with pockets of Evenki speakers living as far west as the Yenisei River. To the southwest, Xibe is spoken in China’s Xinjiang Province, and to the east a number of Tungusic

languages are spoken in the Khabarovsk Krai in Russia's Far East region, and one, Orok, is spoken on Sakhalin Island, albeit by very few people. The prevalence of dialect continua in the Tungusic family, as well as whether the concept of "language" is intended to be a label of genetic relatedness or communicative possibilities, makes it difficult to specify the precise number of Tungusic languages. A common inventory of the currently spoken languages, based loosely on measures of mutual unintelligibility is: Even, Evenki, Manchu, Nanai, Negidal, Oroch (probably now extinct), Orok, Oroqen, Solon, Udihe, Ulcha, and Xibe. Arguably, this list could be expanded or contracted depending on the notion of "language" being used.

The greatest density of Tungusic languages is in the Amur River basin (Nanai, Udihe, Negidal, Orok, Oroch, and Ulcha). Despite this fact, the evidence that can be gleaned from Chinese texts, as well as the oral traditions of these Tungusic peoples, toponyms and aspects of material culture, suggest that the Tungusic *Urheimat* may lie somewhere else. Several different hypotheses have been advanced. Schmidt (1923) and Shirokogorov (1929) both have proposed Manchuria, as Janhunen (2012c) has done more recently; see also Robbeets et al., this volume: Chapter 43. On the other hand, Menges (1968) and Levin (1960) argue for a homeland further to the west in the Lake Baikal region. Pevnov (2012) argues for the Central Amur River basin, though suggests that it was in a mountainous region that may have included parts of Manchuria.

Given the traditional subsistence activities and social organization of Tungusic peoples, many of whom were nomadic or semi-nomadic until the mid-20th century, few of the Tungusic groups ever developed large populations. The exceptions to this are the Manchus and the Jurchens. The Manchus emerged as a powerful force in China in the early 17th century and conquered the Ming empire to found China's final empire, the Qing dynasty. The Jurchens conquered much of

northern China and founded the Jin dynasty in the 12th century. Manchu and Jurchen provide the only Tungusic literary corpora in that predate the 20th century.

The number of Tungusic speakers has never been great, and that is truer today than ever as most ethnically Tungusic people now utilize Russian or Mandarin more frequently than the languages of their ethnic heritage. Indeed, most Tungusic languages are not only endangered, but critically so, including Manchu for which there is only a small number of semi-speakers. Accurate estimates of fluent speakers are difficult to come by, but only three Tungusic languages currently may have more than 10,000 speakers (Solon and Xibe in China, and Evenki in Russia). Most of the other languages number their speakers in the tens or 100's.

6.2 Approaches to Tungusic classification

The concept of a genetic Tungusic unity was proposed in the late 19th century by scholars such as Schrenk (1883) and Grunzel (1895). According to them, the family had a bipartite structure with a Northern and a Southern branch, though this was based as much on geography as linguistic similarity. These early efforts at classification were improved on significantly by Schmidt (1915), who also concluded that there was a basic division between Northern and Southern Tungusic. This bipartite classification was given a much more substantive empirical footing by Cincius (1949) and Benzing (1956), the former of whom also provided sub-groupings within the two major branches. Cincius' classification is given in Table 6.1.

Table 6.1 Classification of the Tungusic languages (Cincius 1949)

SOUTHERN/MANCHU GROUP	NORTHERN/TUNGUS GROUP
Manchu	Evenki

Nanai		-Northern dialect
	-Nanai proper	-Southern dialect
	-Sungari dialect	-Negidal dialect
	-Ulcha dialect	-Solon dialect
	-Orok dialect	
Udihe	Even	
	-Udihe proper	-Eastern dialect
	-Oroch dialect	-Western dialect
		-Arman dialect

Cincius' classificatory scheme, though it remains extremely influential and useful, has been challenged in two significant respects[1]. First, the differences between Manchu and the other members of the Southern Branch are large enough that grouping them together may not be warranted. This was the position taken by Doerfer (1978a), for example, who separated Nanai and Udihe from Manchu and combined them into a distinct "Central" Branch. Table 6.2 depicts this alternative grouping. In order to allow for easy comparison with Table 6.1 above, as well as Tables 6.3–6 below, the "dialects" identified by Cincius are ignored. The close genetic relationship within these clusters of "dialects" is broadly accepted.

Table 6.2 A trinary classification of Tungusic (e.g. Doerfer 1978a)

SOUTHERN GROUP	CENTRAL GROUP	NORTHERN GROUP
Manchu	Nanai	Evenki
	Udihe	Even

While today there is consensus on the need to branch Manchu off from Nanai and Udihe at some level, the second objection to Cincius' classification, her specific placement of Nanai-Udihe within the internal structure of Tungusic, has proven to be more contentious. For example, Avrorin (1960), Poppe (1965), and Menges (1968) see the branching as a subdivision within the Southern Group, rather than a unique branch (Table 6.3). On the other hand, Sunik (1959) and Vovin (1993c) argue that Nanai/Udihe bear more resemblance to the Northern languages than they do to Manchu, and hence place them as a sub-branch of the Northern Group (Table 6.4).

Table 6.3 Nanai-Udihe as Southern Tungusic (e.g. Menges 1968)

SOUTHERN GROUP	NORTHERN GROUP
Manchu	Evenki
Nanai-Udihe	Even

Table 6.4 Nanai-Udihe as Northern Tungusic (e.g. Sunik 1959)

SOUTHERN GROUP	NORTHERN GROUP
Manchu	Evenki
	Even
	Nanai-Udihe

The rather surprising disparity between the positions of Nanai-Udihe in Tables 6.3–4 stems from the fact that while these languages are quite similar to each other, Nanai has some decidedly Southern features, whereas Udihe has Northern ones. Consider two well-known correspondences

that exemplify this point: i) Proto-Tungusic word initial **p-*, which is retained in the Southern Group, is lenited in the Northern Group to *x-*, *h-* or ø; and ii) the front rounded vowel PTg *ü* became a back vowel in the Southern Group and unrounded in the Northern Group.[2] (Both of these correspondences are discussed in more detail in Section 6.4). In both instances, Nanai and Udihe diverge from one another, as shown in Table 6.5, where they are compared with an uncontroversial member of each group.

Table 6.5 Divergent correspondences for Nanai and Udihe

		p-* (pajī* 'grass')	*ü (**xilŋü* 'tongue')
Southern pattern	Nanai	***p**ajaqta*	*ʃiŋm**u***
	Manchu	***f**ojō*	*iləng**u***
Northern pattern	Udihe	***x**aikta*	*iŋ**i***
	Evenki	***h**ajīxta*	*iŋŋ**i***

Due to this contrasting behavior, Ikegami (1974) suggests a more cautious interpretation of the structure of Tungusic classification, one in which there are four primary branches, as in Table 6.6.

Table 6.6 A quaternary classification of Tungusic (e.g. Ikegami 1974)

SOUTHERN GROUP	CENTRAL GROUP I	CENTRAL GROUP II	NORTHERN GROUP
Manchu	Nanai	Udihe	Evenki
			Even

The puzzle of the genetic relationship of Nanai and Udihe has been taken up more recently by Janhunen (2012c) and Georg (2004). In an effort to preserve a classical approach to classification that assumes binary branching, they each, apparently independently, arrive at a sixth possible scheme (Table 6.7), where Cincius' original Southern vs. Northern branches are maintained, but Udihe, along with Oroch, is placed in Northern Tungusic and Nanai, along with Orok and Ulcha, in Southern Tungusic.

Table 6.7 Binary classification with Nanai and Udihe different branches (e.g. Georg 2004)

SOUTHERN GROUP	NORTHERN GROUP
Manchu	Evenki
Nanai	Even
	Udihe

This cursory overview of the history of Tungusic classification is meant to underscore that fairly fundamental questions remain about the genetic relationships among the Tungusic languages. Some of the disparities in how scholars have answered these questions derive from the utilization of different methodologies. For example, Vovin (1993c) develops his classification using lexicostatistics, whereas Doerfer (1978a) determines genetic distances by the overall number of shared isoglosses among languages, whether they be sound correspondences or morphological differences, and Georg (2004) assumes a certain small set of isoglosses to be primary indicators of branch membership. Regardless of the methodology employed, however, it is important to note that the quest to reach a consensus on Tungusic classification is hampered by two well-

known realities. The first is the lack of a literary tradition before the 20th century for any of the Tungusic languages except Manchu (and Jurchen, a now-extinct language that was either a historical precursor to Manchu or a variety of it). This lack of historical data makes it difficult to determine with any certainty if sound correspondences like those in Table 6.1 are a result of shared innovations among languages, or whether the sound changes occurred independently, or whether language contact is the key variable. Moreover, nearly all of the Tungusic correspondences result from highly common sound changes (e.g. lenition of consonants in between vowels, loss of final vowels, and the like), so the odds of them occurring multiple times within the family are not low. Indeed, all classifications of Tungusic assume that some of the isoglosses are only of secondary importance in determining the cladistic structure of the family.

A second challenge for Tungusic classification, particularly when comparing morphological and syntactic properties, comes from the fact that all of these languages have been noticeably impacted by contact with non-Tungusic languages, and, quite likely, by each other as well, though the inter-familial contact is harder to establish. Ikegami (1979), for example, argues that many of the grammatical properties, which set Manchu apart from other Tungusic languages, such as its third person pronouns (*i* 3SG / *ce* 3PL vs. the typical Tungusic root *nungan*) or the zero ending in the imperative, are best explained by Mongolic influence. Similarly, Baek (2016) proposes that differences found among Tungusic languages in how they mark the 3rd person on finite verb forms, how they use converbs formed by *-mi*, and other syntactic features result from contact with distinct sets of languages, such that Tungusic languages can be categorized by their geographical location into one of three linguistic areas: Siberia, the Russian Far East and China (the last of which was also observed by Tsumagari 1997). If this general picture is accurate, then morpho-syntactic comparisons may more often be identifiers of areal relatedness than genetic

relatedness. However, even the comparative work that restricts itself to sound correspondences commonly hypothesizes that some of them are contact-induced (e.g. Georg 2004 suggests that the loss of *-*g*- in both Nanai and Udihe is due to contact and not genetic relatedness). Certain linguistic similarities, therefore, may be as much a product of geography as genetic familial relationship.

As a consequence of these facts, some scholars have concluded that a traditional binary branching tree structure for Tungusic is, minimally, not possible to determine (as implied by the trinary branching hypothesis in Table 6.2 or the quaternary hypothesis in Table 6.6), or, possibly, that nearly any branching scheme is elusive because of so many instances of contact-induced historical change. This is, for instance, the conclusion of Whaley et al. (1999: 313), who write:

> The geography of this region and the traditional social organization of both Tungusic peoples and their linguistic neighbors (Turkic, Mongolic, Indo-European, and more recently Sino-Tibetan speakers) have long encouraged large-scale multilingualism and cultural intercourse, classic conditions for the formation of convergence areas. Morphological, and presumably phonological, properties, even those of an unusual sort, can be passed from language to language (see Li and Whaley 1998). All this places a special burden on any claims that linguistic similarities are indicative of genetic relationships.

The lack of historical data and the pervasiveness of contact-induced language change are significant limitations on the confidence we can have in reconstructing binary-branching relationships in the Tungusic family. That said, in the next section, we will lay out the evidence

for a conventional tree-structure interpretation of the facts. Before doing so, it is useful to summarize points of agreement about the Tungusic language family.

(i) The Even and Evenki language/dialect complexes bear close family resemblance but are also different enough to warrant being identified as sub-branches within the Northern Tungusic Group.

(ii) The Manchu complex (Manchu, Jurchen and Xibe) has characteristics not shared with other members of Tungusic that require explanation, either in that the complex is its own unique branch within the family (contra Cincius 1949), or in that it bears a close genetic relationship to the Nanai complex (and possibly also the Udihe complex), but this relationship has been obscured by Manchu's change resulting from Manchu's contact with Mongolic and Sinitic languages.

(iii) The relationship between the Nanai complex and the Udihe complex, both to each other and to the family as a whole, presents the single biggest challenge to developing an accurate Tungusic family tree.

(iv) Different methodological approaches and assumptions will lead to different conclusions about the validity of different classifications, or even the possibility of developing a valid classification.

6.3 Proto-Tungusic segmental inventories

While the classification of Tungusic has eluded consensus, a reconstruction of the sounds in the proto-language is a simpler matter. Perhaps this is not much of a surprise given the relatively shallow time-depth of the family. The consonant inventory is shown in Table 6.8.

Table 6.8 Consonant inventory of Proto-Tungusic (Benzing 1955a)

	labial	dental	palatal	velar
voiceless stop	*p	*t		*k
voiced stop	*b	*d		*g
voiceless affricate			*č	
voiced affricate			*ǯ	
voiceless fricative		*s		*x
nasal	*m	*n	*ń	*ŋ
liquid		*l, *r		
glide	*w		*j	

It should be noted that Cincius (1949) reconstructs a larger set of palatal consonants (ḿ, d´, and ś) than that given in Table 6.5. However, Cincius' additional palatals can be equally well-accounted for by the presence of *i*-diphthongs in the proto-language. For example, the Manchu word *ḿali-* 'to measure' is consistent with either **ḿali-* or **miali-*. Most Tungusologists opt for the simpler consonant inventory.

There is less agreement about the details of the vowel inventory of Proto-Tungusic. It is generally accepted that the vowel system involved three vowel heights and a front vs. central vs. back distinction. Benzing (1955a) and others have proposed a system of eight vowels (though see Ko et al. 2014 who argue for a seven-vowel system, see also Joseph et al., this volume: Chapter 29), which is summarized in Table 6.9.

Table 6.9 Vowel inventory of Proto-Tungusic

	front	central	back
high	*i, ü		*u
mid	*e	*ö	*o
low		*a	

It is likely that Proto-Tungusic had distinctive vowel length as well, since this is a characteristic feature of most languages in the family. It is also likely that the proto-language had a vowel harmony system, which is, again, a characteristic feature of most Tungusic languages. Because the details of the vowel harmony systems in modern Tungusic languages vary, reconstructing the original harmony processes is challenging, though Ko et al. (2014) make a good case for RTR harmony.

6.4 A binary-branching classification of the Tungusic languages

With the introduction to Tungusic classification and a summary of basic phonological details in place, we are now in a position to describe those features that offer the most promising evidence for sorting out the genetic relationships in the family. In this section, we restrict the evidence to sound correspondences under the assumption that, as a general rule, morphosyntax is more susceptible to borrowing than phonology. By restricting attention to sound correspondences, then, we can reduce the proportion of contact induced historical changes.

6.4.1 *t-

Drawing on an observation made by Benzing (1956) that there is a small set of lexemes in which Manchu word initial *s*- corresponds with *t*- in other Tungusic languages, Norman (1977) identifies the condition under which the sound change **t* > *s* occurred in Manchu, namely when it was followed by **j of *ń* in the same stem, see (1) and (2).

(1) PTg **teja* 'meat cooked in its own juices' > *sija* (Manchu), *tijaki* (Evenki), *tigje* (Negidal), *cijiki* (Even), *čeoki* (Nanai), *tija* (Oroch)

(2) PTg **tuńga* 'five' > *sunǯa* (Manchu), *tunŋa* (Evenki), *toŋnŋa* (Negidal), *tunŋin* (Even), *tojŋga* (Nanai), *tuŋa* (Udighe)

Note that this correspondence, though suggestive of an early split between Manchu and the other Tungusic languages, is compatible with the view that Nanai-Udihe are also in the Southern Branch (Table 6.3), and with the view that Nanai is in the Southern Branch, while Udihe is in the Northern Branch (Table 6.6). To maintain either of these views, however, one would have to identify some other early innovation that divided the two groups, and then state that within the Southern Branch, the sound change **t>s* took place differentiating Manchu from Nanai and Udihe. It should be noted that the word for 'five' in Jurchen is *ĉunǯa* and it is *sunja* in Xibe, which is also consistent with the notion that within the Southern Branch, one of the early innovations was the weakening of word initial **t* in the Manchu sub-branch within the Southern Group.

6.4.2 *-b-

Intervocalic PTg *-*b*- evinces a regular set of correspondence in members of the family, where Manchu retains the proto-sound and the other Tungusic languages have different degrees of weakening, as in (3) and (4).

(3) PTg **teb*- 'to place, put' > *tebu*- (Manchu), *tew*- (Evenki/Even), *teu*- (Nanai/Udihe)

(4) PTg **sube*- 'end, edge, top' > *subexe* (Manchu), *suwerē* (Evenki), *hūre* (Even), *suwe/sue* (Nanai), *sue* (Udihe)

The simplest explanation for these data is that the lenition of *-b- is an early innovation that sets the Northern Branch apart from Manchu. Alternatively, if Nanai (and possibly also Udihe) is in the Southern Branch, then there was a lenition process that occurred independently of the one that occurred in the Northern Branch.

6.4.3 *-k-

Intervocalic PTg *-*k*- weakens to varying degrees in Tungusic languages other than Even (as well as some Evenki dialects) (see 5 and 6).

(5) PTg **baka*- 'to find, obtain' > *baχa*- (Manchu), *baka*- (Evenki), *baq*- (Even), *bā*- (Nanai), *ba*- (Udihe)

(6) PTg **šikēn* 'urine' > *sike* (Manchu), *čikēn* (Evenki), *čikin* (Even), *čiẽ* (Nanai), *čea*- 'urinate' (Udihe)

Regardless of how one proposes to classify Tungusic, it is necessary to posit several independent processes of -*k*- lenition. At the deepest level of branching in the family, the correspondence suggests that the Southern Branch had -*k*- lenition (at least in some environments, cf. 5 and 6), while the Northern Branch did not. If the Nanai (and possibly Udihe) cluster of languages is part of the Southern Branch, then it underwent a further process of weakening such that *-*k*->ø, an innovation that sets it apart from the Manchu cluster of languages. On the other hand, if it is a member of the Northern Branch, then Nanai/Udihe together branched off from Even/Evenki with the lenition if -*k*-. It should be noted that languages closely related to Evenki also have a lenition of the -*k*-, cf. *baka*- (Evenki) with *baxa*- (Negidal, Solon, Oroqen), which suggests yet another independent lenition of *$k > x$.

6.4.4 *-g-

Reflexes of intervocalic -*g*- are quite similar in their distribution to those of intervocalic -*b*- (Section 6.4.2), and the two isoglosses could be seen as reflexes of the same sound change. As was the case with -*b*-, the lenition process can be taken as an innovation of the Southern Branch, and then, independently, lenition occurs within the Northern Branch for Udihe and Nanai if they are in the Northern group (see 7 and 8).

(7) PTg *_togo_ 'fire' > _tuwa_ (Manchu), _toɣo_ (Evenki), _toɣ_ (Even), _tao_ (Nanai), _tō_ (Udihe)

(8) PTg *_tuge_ 'winter' >_tuweri_ (Manchu), _tuɣe_ (Evenki), _tuɣəni_ (Even), _tue_ (Nanai), _tue_ (Udihe)

At least on a superficial level, the similarity between Nanai and Udihe with respect to intervocalic *-b-* and *-g-* appears to be good evidence for the placement in the same branch. However, that evidence is not as strong as it first seems. Languages related to Nanai do not undergo a full loss of *-g-*, as does Nanai, e.g. **togo* 'fire' > *tawa* (Orok and Ulcha), and **tuge* > *tawe* (Orok) and *tawa* (Ulcha). For that reason, one could also interpret the data to indicate that the Southern Branch underwent a lenition process of **-g-* > *w* and that, then, within the Southern Branch, Nanai (and Udihe if it is Southern) underwent a second lenition of **w* > ø. Similar facts would be true for the Northern Branch, in which the **-g-* weakens to *-ɣ-*, and then Udihe/Oroch underwent further weakening of **-ɣ* > *-w* > ø (the intermediate step is suggested by some dialects of Oroch in which *-w-* is maintained).

Though not problematic for the larger classification picture being described here, there are some intriguing comparative differences within the languages related to Evenki. In the correspondences, Solon has *-g-* (cf. **togo* 'fire' > *tog, togo*) and Oroqen has undergone a loss of *-g-* (cf. **togo* 'fire' > *tō*).

6.4.5 *p-

The reflexes of PTg **p-* may be the single most significant piece of evidence for the dendritic structure of the family, as in (9), (10) and (11).

(9) PTg **palŋa* 'palm of hand' > *falaŋgu* (Manchu), *hanŋa* (Evenki), *hanŋi* (Even), *pajŋa* (Nanai), *xaŋa* (Udihe).

(10) PTg **peli-* ‘to walk’ > *fele-* (Manchu), *helde-* (Evenki), *höl-* (Even), *pul-si-* (Nanai), *xuli-* (Udihe)

(11) PTg **pulńe-* ‘ashes’ > *fuleŋgi* (Manchu), *hulteptēn* (Evenki), *hultēn* (Even), *puńektē* (Nanai), *xulepte* (Udihe).

As Georg (2004) argues, the simplest way to account for these facts is that a Southern Branch that includes Nanai maintained the PTg **p-*, while in the Northern Branch **p* > **x*. In the Southern Branch, the Manchu-related dialects weakened **p* > *f*. In the Northern Branch, **x-* > *h-* for Evenki/Even (though not in Negidal, where *x-* is found), while **x-* was retained in Udihe-related languages. There is also a further *h-* > ø in Oroqen and Solon (**pulńe* > *ulepten*).

That, of course, is not the only way to interpret the data. Alternatively, one could take **p-> f-* as a characteristic of the Southern Branch. The Northern Branch (including Nanai) maintains **p-*, and within the Northern Branch, Evenki/Even undergo **p* > *x*. Within the Evenki/Even languages, there is further *x* > *h* > ø. Within the Nanai/Udihe languages, Nanai maintains **p-* and Udihe shows further weakening, i.e. **p-* > * *x-*.

6.4.6 *-i/-u

The loss vs. maintenance of stem final high vowels is occasionally mentioned as a useful isogloss for establishing a Northern vs. Southern Branch in Tungusic, typically with **dili* ‘head’ (12) used to exemplify the correspondences.

(12) **dili* ‘head’ > *ʒili* (Manchu), *dil* (Evenki), *dil* (Even), *ʒili* (Nanai), *dili* (Udihe)

If this, indeed, were a primary isogloss, it would provide a convincing piece of evidence that Nanai and Udihe are both in the Southern Branch (as Cincius 1949 proposed). However, at least according to Doerfer (1978a), the loss of stem final high vowels has been known to raise more questions than it answers. Most problematically, Evenki dialects are not uniform, nor are the languages related to Evenki. Oroqen and Solon preserve the stem-final high vowel in most lexemes, whereas Negidal does not (cf. the demonstrative 'that': *tari* (Oroqen), *tari/tayi* (Solon), *tar* (most Evenki), *tay* (Negidal)). An interpretation of this isogloss that is most consistent with the data is that the loss of stem-final *-i* is only of value in making sense of the genetic relationships within the Evenki cluster of languages, but not of importance for deeper branching in the language family.

6.4.7 *ü$_1$

Another long-recognized distinction between Southern and Northern Tungusic are the reflexes of **ü* in initial syllables, see (13)–(15).

(13) **tügde* 'rain' > *tuxe-* fall (Manchu), *tigde* (Evenki), *tīd* (Even), *tugde* (Nanai), *tigde* (Udihe)

(14) **üse-* 'to grow' > *use* 'seed' (Manchu), *isew-* (Evenki), *isu-* (Even), *use* 'seed' (Nanai), *jehu-* (Udihe)

(15) **lümŋe-* 'to swallow' > *nuŋgi-* (Manchu), *nimŋe-* (Evenki), *ńimŋi-* (Even), *luŋbe-* (Nanai), *niŋme-* (Udihe).

Notice that for this feature, Nanai (and the languages related to it) patterns with the Southern Branch and Udihe (and the languages related to it) with the Northern Branch. If this sound change is taken to be a primary isogloss, this is good evidence for the classification scheme in Table 6.7. The alternative is to have two independent changes of *ü* > *u*, one for Manchu, and one within the Northern Branch that indicates either an early break between Nanai and the other languages, or one of the innovations that sets Nanai apart from Udihe.

6.4.8 *u_2

Both Doerfer (1978a) and Georg (2004) identify a *u* ~ *i* correspondence in the second syllable of a word as characteristic that distinguishes the Northern Branch (16).

(16) a. **aduli* 'net' > *adil* (Evenki), *adili* (Udihe). Cf. *aduli* (Ulcha) (Georg 2004)

b. **lelu* 'gusset' > *leleke* (Negidal), *lelū* (Nanai), *leli* (Udihe)

However, the value of this isogloss is highly suspect. First, in the overwhelming number of instances in which there is a second syllable **u*, the *u* is maintained in all the daughter languages. Second, there are few actual cognate sets that allow for a comparison of Manchu, Evenki (or a related language), Even, Nanai (or a related language) and Udihe (or a related language). Third, undisputed Northern languages often behave differently from one another when there is variation. For example, PTg **maltu*- 'to bend' becomes *maltu*- in Evenki, but *malti*- in Negidal. Finally, there are instances in which Northern languages maintain u_2, but a member of the Transitional Branch (according to Doerfer 1978a) or Southern Branch (according to Georg 2004)

has *i*, as in **dulbu* ‘stupid’ > *dulbun* (Evenki), *dulbi* (Nanai). Therefore, the value of this piece of evidence is highly suspect.

6.4.9 Summary

In working through the sound correspondences that are most often used to classify Tungusic languages, we have noted that two of them are quite problematic: the loss of a final high vowel and $u_2 > i$. Table 6.10 summarizes the isoglosses for the remaining six features. The shaded portion of each column identifies a possible shared innovation within the family for each feature.

Table 6.10 Tungusic isoglosses

	*t-	*-b-	*-k-	*-g-	*-p-	ü1
Manchu	s	b	k/χ	w	f	u
Nanai	t/č	w/ø	ø	ø	p	u
Udihe	t	ø	ø	ø	x	i
Evenki	t	w	k	k	h	i
Even	t/č	w	k/q	k/ɣ	h	i

6.5 Interpretation

The set of sound changes discussed in the previous section do not lead to a definitive picture of the classification of Tungusic. They do, however, suggest that one of three scenarios is a likely way to conceptualize the genetic relationships among the languages in a classical comparative approach, at least based on sound correspondences. The first is to posit a Southern Branch consisting of Manchu, Xibe and Jurchen on the basis of a **t* > *s* innovation in the Southern

Branch and a *-*b*- > *w* innovation in the Northern Branch (Table 6.4). Under this interpretation, intervocalic velar deletion (i.e. -*k*- > ø and -*g*- > ø) is an innovation that sets Nanai-Udihe apart from the other Northern Tungusic languages. If so, then **p* and **ü*$_1$ require some explanation since Nanai and Udihe do not pattern together. The second scenario is that Nanai is in the Southern Branch on the basis of the Northern languages shared innovations *ü* > *i* and **p* > *x/h* (Table 6.7). This scenario requires a larger number of independent lenition processes (for -*b*-, -*k*- and -*g*-), though none of them are unusual historical processes, and following Georg (2004), heavy contact influence between Nanai and Udihe can be invoked as part of the explanation. With regard to the third scenario, proposed by Cincius (1949), Nanai and Udihe are both in the Southern Branch, which is not altogether implausible, but it does require the greatest number of additional changes (and a reliance on the language-contact argument for linguistic similarities) to account for Udihe's regular correspondence with the Northern languages with **t*-, *-*b*-, **p*- and *ü*. From all this, it becomes clear why the trinary branching (Table 6.2) and quaternary branching (Table 6.6) schemes hold some appeal since they do not commit one to an alternative scenario with limited evidence.

Examining the sound correspondence data has led to the three most likely classifications for Tungusic using a binary-branching approach. Short of discovering additional isoglosses that tip the scales towards one or the other, the next obvious step is to admit additional sorts of data. Morphological similarities/differences could be examined. Indeed, they have regularly been included in most efforts to classify Tungusic languages. Robbeets (2015) restricts her focus of study to bound morphology since they have proven to be a better diagnostic for shared retentions than morphology that encodes morpho-syntactic features. Using this method, she argues for the first scheme mentioned above: there is Southern branch that includes Manchu, Xibe and Jurchen,

and a Tungusic branch. The Tungusic branch splits into a Southern group (Udihe and Nanai) and a Northern group (Even and Evenki).

However, when we examine other types of morphology we readily admit that many similarities are as likely to be areal as genetic phenomena. For example, the fact that the undisputed Southern languages (Manchu, Jurchen and Xibe) have fewer inflections than Nanai-Udihe is as likely a result of extended contact with Sinitic than serving as an innovation that serves as a reliable marker of genetic distance.

Much morphological evidence also offers conflicting evidence. For example, Georg (2004) points to the existence of a prosecutive case suffix *-dulī* that occurs in Udihe, Evenki and Even, but not in Nanai and Manchu. He makes the reasonable suggestion that *-dulī* is an innovation of the Northern Branch, which if true, supports his proposal that Nanai, but not Udihe, is a Southern language. However, the isogloss is not clear-cut since there are a number of dialects in Northern Tungusic that do not have it.

The imperative system of Tungusic has commonly been identified as an area of comparison that might shed light on genetic relations within the family. The existence of bare stem imperatives is unique to the Manchu cluster of languages, for example, though yet again this could have been induced by contact with Sinitic and Mongolic. Fuente (2012), however, presents a case that Manchu adheres more closely to the Proto-Tungusic imperative paradigms and that Northern Tungusic (including Nanai) developed an imperative system with person endings. If he were correct, this would be the sort of morphological evidence that would favor placing Nanai (and Udihe) into the Northern Branch, since it is not as easily accounted for as a contact-phenomenon.

While any of the three binary branching schemes discussed here is plausible, the first (the Manchu complex forms the Southern Branch, whereas all other Tungusic languages are in the Northern Branch) requires fewer independent sound changes than the other two. This, combined with Robbeets' proposal based on bound morphology and Fuente's proposal about imperative morphology, may tip the scales slightly in favor of the first scheme. However, we stress again that the evidence is not definitive. In the next section, we examine a different kind of analysis for the classification of Tungusic to determine whether this sheds additional light on the best binary classification.

6.6 A Bayesian phylogenetic approach to Tungusic classification

Another type of data that tends to be relatively more resistant to borrowing is basic vocabulary, at least more resistant than most kinds of morphosyntactic features. Some of the classifications discussed in Section 6.2 were based on basic lexicon (sometimes alongside with other features). For example, Vovin (1993c) developed his classification applying the lexicostatistic method to the Swadesh 100 word list of the major Tungusic languages. Whaley et al. (1999) also applied the lexicostatistic method to a 200 word list of the Northwest Tungusic languages, i.e. Evenki, Negidal, Solon and Oroqen. In this section we present an attempt to develop a Tungusic classification with the use of the Bayesian phylogenetic method. The question arises whether this classification will support one of the earlier proposed classifications presented in Section 6.2 or it will form a new one. The internal structure of 4 Tungusic clades (Manchu, Nanai, Udihe and Northern) is almost universally accepted and, therefore, can serve as a test for the reliability of the Bayesian phylogenetic classification.

6.6.1 The Bayesian method

A Bayesian phylogenetic approach has been introduced to linguistics from the field of Evolutionary Biology, and, over the last several decades, it has been applied to the classifications of language families (Gray et al. 2009; Kitchen et al. 2009 and others). The Bayesian phylogenetic method allows to determine the internal structure of a language family, estimate the robustness (likeliness) of a classification and determine the approximate time-depth of splits between separate branches or languages using known prior historical information. The general idea is that a genealogical language tree is built according to a chosen model and set of parameters. The models differ in rates given to the loss or innovation of an item (a cognate with a specified meaning, in the case of basic vocabulary). Three basic models (Continuous Time Markov Chain (CTMC), Binary Covarion and Stochastic Dollo) have been applied to the Tungusic data. The CTMC model assumes that items can be gained and lost at the same rate. The Binary Covarion Model allows each item to have different rates along different branches. The Stochastic Dollo model posits that each item can be gained only once in the tree, but it can be lost multiple times. These three models were then combined with the Gamma Model and clock models. The Gamma Model allows different items to have different rates along the whole tree. Clock models reflect the change of rates along the whole tree. A strict clock model assumes that rates of change are constant while relaxed clock models are more realistic and imply the variation of rates across the tree. Computational methods allow us to build millions of trees with the use of chosen models and parameters. After the generation of possible trees, a maximum clade credibility tree is selected, i.e. a tree which fits best with the chosen parameters. The calculation of the Bayes factor comprises the comparison of likelihood scores of different analyses (i.e. the probability for the data to evolve according to the given model). The Bayes

factor allows for the selection of an analysis with a particular model which fits best with the observed data.

6.6.2 Tungusic data

We collected basic vocabulary items on the basis of Leipzig-Jakarta-Jena list. This list includes 254 items—words whose meanings are resistant to borrowing. The first 100 items are taken from the Leipzig-Jakarta list (Haspelmath and Tadmor 2009a) and ranked from the most stable to the least stable items. The rest of the items are words drawn from the Swadesh 200-word list, but which were not included into the Leipzig-Jakarta list. Basic vocabulary lists were collected for 18 Tungusic languages and transitional "doculects", i.e. varieties with an unclear status of a language or a dialect. All the data were collected from dictionaries, wordlists, grammar descriptions and, in case of gaps, from texts. (NB: Orok data were drawn from the fieldnotes of Patryk Czerwinski). The analyzed languages and "doculects" are as follows: Hezhe (or Hezhen, Kilen), Oroch, Udihe, Standard Nanai (Najkhin dialect), Kur-Urmi (or Kili), Ussuri Nanai (Bikin dialect), Orok (or Uilta), Ulch, Jurchen, Manchu, Xibe, the Olsky dialect of Even, the Momsky dialect of Even, Solon, Evenki, Negidal, Oroqen, and Khamnigan Evenki. Although dialects of the same language were not generally taken into consideration, two dialects of Even were included because an approximate time of their split helps to measure the time depth of other nodes of the tree.

In some cases more than one lexeme was chosen for an item of the basic vocabulary list, and it was not possible to get all items for every language, especially Jurchen, which lacks 72 of 254 items. Nevertheless, the Jurchen data were taken into account as they did not lead to unexpected results and helped to specify the time depth of Tungusic branches. Languages were compared to

each other on the basis of cognates. Cognate lexemes were determined according to Starostin et al. (2003) and Cincius (1975, 1977). A small number of evident loanwords was excluded, although this did not affect the results of analysis. Borrowings from other Tungusic languages were not excluded for Kur-Urmi and Hezhe because the nature of their substrate is still questioned.

The data were arranged in a binary alignment where "1" refers to the existence of a cognate with a particular meaning in a particular language and "0" refers to the absence of the corresponding cognate. The Table 6.11 demonstrates a part of the observed alignment before its transformation to the binary format and after it.

Table 6.11 Fragment of the Tungusic alignment

	*ńūri-kte 'hair'	*puńe- 'hair'	*xü:kte 'tooth'
Manchu	—	funijəhə	wəihə
Nanai Najkhin	nuktə	—	huktə
Evenki	ńuriktə	—	iktə

	*ńūri-kte 'hair'	*puńe- 'hair'	*xü:kte 'tooth'
Manchu	0	1	1
Nanai Najkhin	1	0	1

Evenki	1	0	1

6.6.3 Analysis

The data were analyzed in the program BEAST 2.4.7 (Bouckaert et al. 2014) and Tracer 1.6.0 (Rambaut et al. 2014). Four different models from the package “Babel” were used: Continuous Time Markov Chain (CTMC), Binary Covarion, Stochastic Dollo, and Pseudo Dollo. Both strict and relaxed clock options were chosen for all of them. Unfortunately, at the moment we have not reached the final reliable results for all the analyses so we present here one of the best current results which can change later during the further work. Figure 6.1 demonstrates the tree structure of the Tungusic languages (a maximum clade credibility tree, which was generated with the program TreeAnnotator 1.4.3). The numbers on the nodes denote posterior probability.

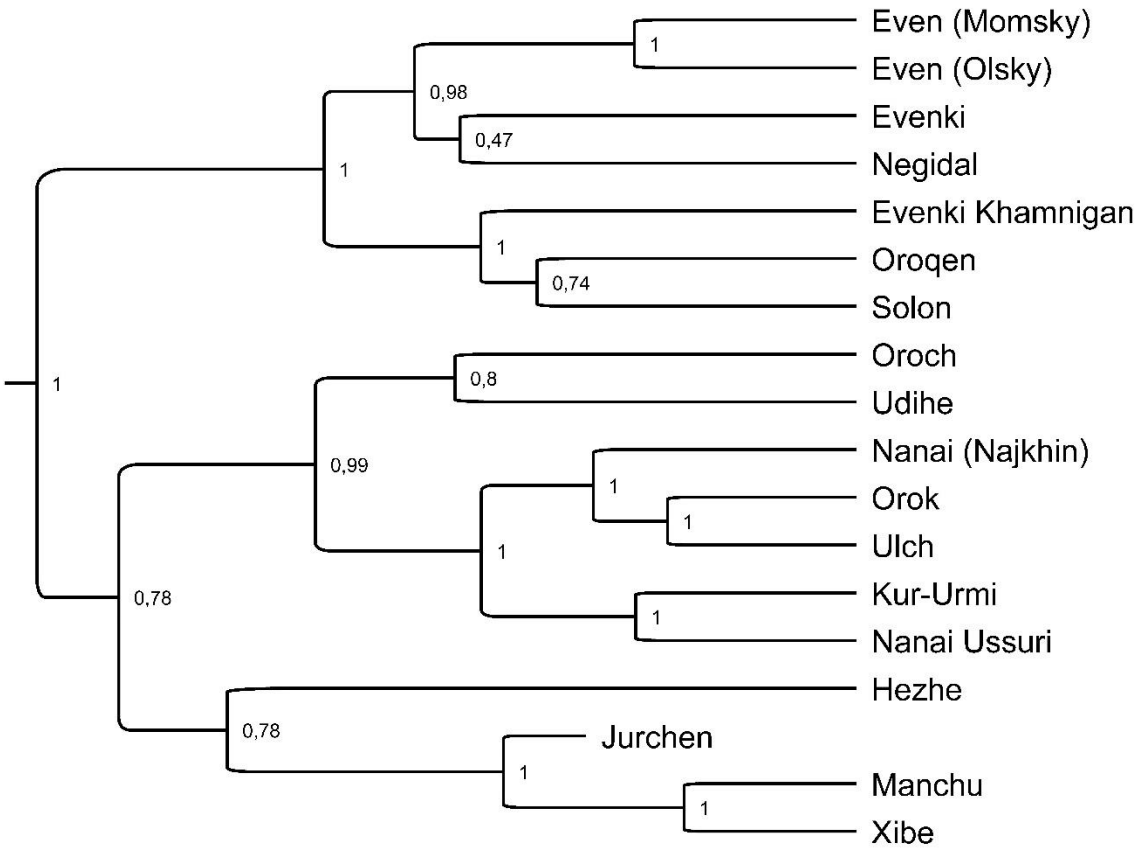

Figure 6.1 Maximum clade credibility tree of the Tungusic languages. Pseudo Dollo, relaxed clock models. Chain length 100 000 000.

It should be noted that BEAST builds trees only with bipartite structure, i.e. each node can be split only in two branches rather than three or four. However, nodes that appear to be tripartite according to the observed data can be detected by a very short distance between two nodes and a very low posterior probability of the shallower node (see the structure of the branch including Even dialects, Evenki and Negidal).

For time calibration, three approximate dates were selected. First, we chose the disappearance of Jurchen texts and the appearance of Manchu texts. As was mentioned earlier, it is unclear whether Manchu is a descendent language of Jurchen or a later

attestation of a Jurchen dialect. Either way, this period of textual transition provides a helpful reference point. Jurchen writing ceased in the 13th century, and the first known Manchu manuscript is dated to 1599 (Gorelova 2002: 50). Therefore, Jurchen is no longer attested and Manchu is attested 351 years before present[3]. Second, the historical record indicates that the Xibe people (who were a Manchu tribe) were resettled to the northwest of China (a non-Manchu speaking area) in 1764 (Gorelova 2002: 31). Therefore, this date can be used as a point of a split into the distinct languages of Xibe and Manchu. We chose the third date for calibration from the earliest mention of the reindeer-breeding Tungusic people in the historical record (presumably Evens) near the Okhotsk sea in 1655 (Turaev 1997: 51–52). We assume that this is an approximate point of divergence of the Even dialects.

The calibration was performed with the package "Sampled ancestors" using the Fossilized Birth Death model (Gavryushkina et al. 2014). The probable evolution of the Tungusic family across time is presented in Figure 6.2. Dates are given before present, i.e. before 1950. The bars show a time period, in which a specific split has taken place with a probability of 95%.

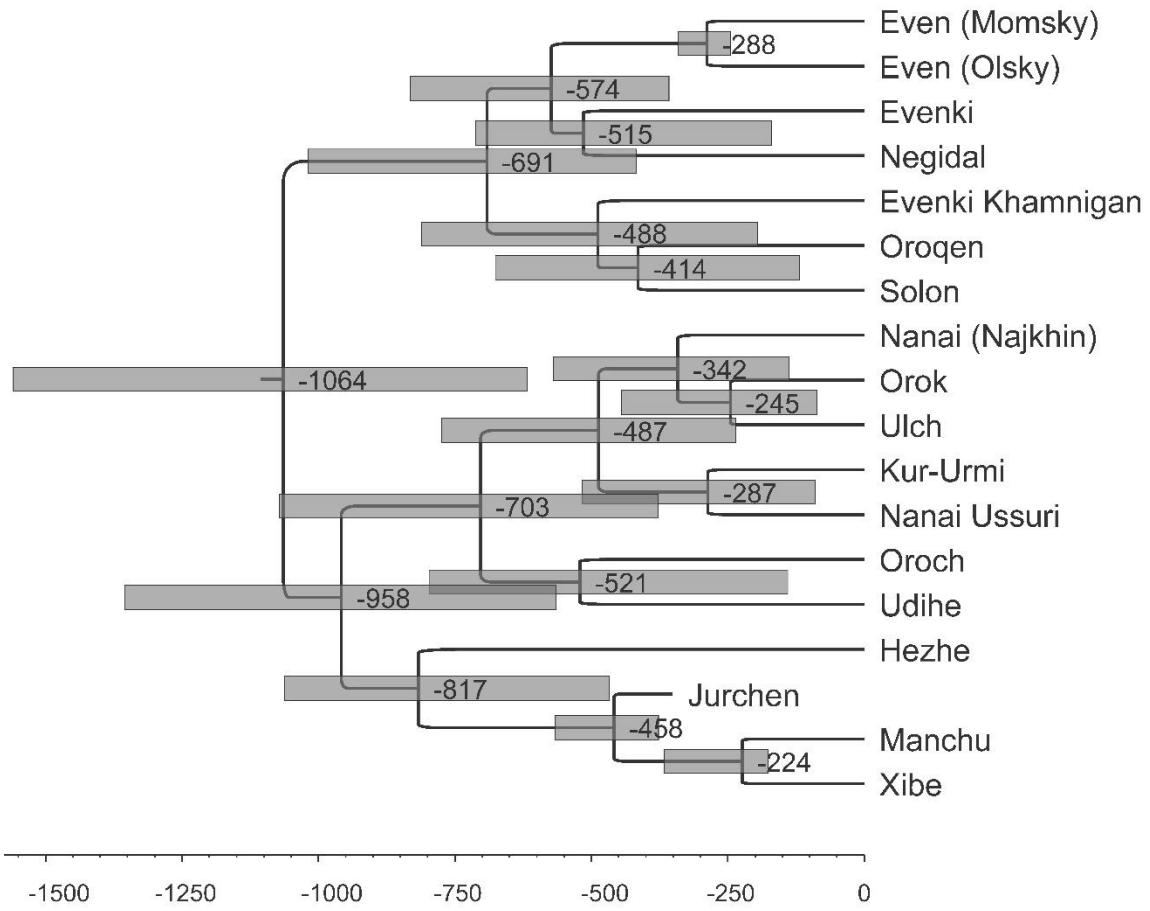

Figure 6.2 Time depth of the Tungusic family[4]

6.7 Discussion and summary

Four major Tungusic complexes are revealed by the phylogenetic methods: a Northern complex (Even dialects, Evenki, Negidal, Khamnigan Evenki, Oroqen and Solon), a Nanai complex (Kur-Urmi, Ussuri Nanai, Najkhin Nanai, Ulch, Orok), an Udihe complex (Udihe and Oroch) and a Manchu complex (Jurchen, Manchu and Xibe). The results of the Bayesian methods align well with previous comparative work, in which there is broad consensus about these four complexes. In addition, the most common view of the Northern Branch, namely that there is an Even complex that forms a distinct sub-branch from Evenki and Negidal, is supported by the Bayesian analysis.

One unexpected result is the position of the Hezhe language in the tree. As can be seen in Figure 6.1, Hezhe is represented as splitting off from the Manchu complex as a separate branch. However, based on the historical record, oral tradition and other comparative analyses, scholars have been uniform in proposing that Hezhe is a variety of Nanai that has been heavily influenced by Mandarin and other Chinese Tungusic languages. Presumably, then, the unanticipated result of the Bayesian analysis must be explained in terms of language contact.

The tree in Figures 6.1–2 supports the hypothesis that the Southern subgroup includes Manchu, Nanai and Udihe complexes while the Northern Tungusic languages form a separate subgroup. As for the Udihe and Oroch languages, they form a separate subgroup together with the Nanai complex. This structure corresponds to the classification proposed by Cincius (1949) and Benzing (1955a), see Table 6.1. But the distance between two first nodes is quite short and the posterior probability is not very high—0,78—implying that these two nodes may change their places. So, our Bayesian analysis does not exclude the triple classification proposed by Sunik (1959) and others, in which the Manchuric branch separated first, see Table 6.4. Thus, it can provide additional evidence for our conclusion in Section 6.5.

Based on the three time-correspondences used to date the genetic diversification of the family, the first split, between Manchuric and other Tungusic languages, goes back to the 9th century AD or a time period of 5–14 centuries AD. This date is more recent than that proposed by other scholars. Both Janhunen (2005b) and Pevnov (2012) suggest the South-North split to have taken place about 2000 years ago. Robbeets (2015) suggests 220 AD for the earliest divergence (see also Robbeets et al., this volume: Chapter 43).

Still, a later date beginning with the 5th century AD does not contradict any available facts and seems equally plausible.

The results obtained with the Bayesian methods should, of course, also be considered as one hypothesis of Tungusic classification among others developed with different methods (quantitative or classical comparative). The analysis here, like all those before it, depends on the methods being applied, the data and on the technical parameters. Moreover, the Bayesian methods do not take into account processes of convergence: e.g., there is no guarantee that Udihe and Nanai complexes form a single subgroup due to shared innovations rather than areal contacts. However, it is striking that, when binary branching is presupposed, the results generated by Bayesian methods support the Southern-Northern analysis and do not exclude the simplest Manchuric-Tungusic analysis using the comparative method. The Tungusic family is best represented as having a Southern Branch that consists of the Manchu and Udihe-Nanai complexes and a Northern Branch comprising Even-Evenki branches, while the Manchuric-Tungusic analysis where the Manchu complex separated the first and then the Tungusic branch split into Even-Evenki and Udihe-Nanai complexes is also possible.

Acknowledgments

We are grateful to Czerwinski for his Orok fieldnotes and to Koile for his help with the Bayesian analysis. The research leading to the results presented in Section 6.6 has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 646612) granted to Martine Robbeets.

[1] As a result, the language family is frequently referred to as “Manchu-Tungusic” or “Tunguso-Manchu”. For this chapter, “Tungusic” is used as the name of the language family.

[2] Recent work (e.g. Ko et al. 2104) argues that the **ü > u ~ i* correspondence is conditioned by a vowel harmony system in PTg. This has no bearing on the point being made here.

[3] “Before present” is traditionally calculated before 1950.

[4] The technical parameters are as follows (dates are given in centuries before present). Tip Date for Jurchen 3.51 BP. Priors: 1) Jurchen, Manchu, Xibe: Log Normal, M = 0.1, S = 1.6, Offset = 3.48, Mean in Real Space, monophyletic; 2) Manchu, Xibe: Log Normal, M = 0.2, S = 1.2, Offset = 1.76, Mean in Real Space, monophyletic; 3) Momsky Even, Olsky Even: Log Normal, M = 1.0, S = 0.3, Offset = 2, Mean in Real Space, monophyletic.

This is a draft version of a chapter that appears in Robbeets, M. and A. Savelyev (eds). *The Oxford Guide to the Transeurasian Languages* (OUP, 2020), see https://global.oup.com/academic/product/the-oxford-guide-to-the-transeurasian-languages-9780198804628. The research leading to these results has received funding from the European Research Council under the Horizon 2020 Program/ ERC Grant Agreement n. 646612 granted to Martine Robbeets.