The Origins of the Korean Language: A Linguist's Perspective

J. Marshall Unger
The Ohio State University

Abstract

On the basis of the comparative method, developed over more than two centuries of empirical study, the best results to date are that the present-day Korean and Japanese languages had a common source, called proto-Korean-Japanese. Korean and Japanese are more similar to one another than either is to any of the languages spoken in adjacent parts of Asia. That is as far as pure linguistics takes us at present. Other scientific disciplines must be utilized to determine when and where proto-Korean-Japanese was spoken, when its speakers separated into pre-Korean and pre-Japanese groups, and when the descendants of those groups resumed contact on the Korean peninsula prior to the migration of most pre-Japanese speakers to Japan.

Keywords: comparative method, proto-Korean-Japanese, historical linguistics

Introduction

Academic disciplines such as history, archaeology, physical anthropology, and human genetics can all shed some light on the origins of a particular language, such as Korean, but it is in principle impossible to prove scientifically that two languages are unrelated. Theoretically, the possibility that all human languages evolved from one enormously old prehistoric language cannot be excluded. Therefore, if one wishes to know what the origins of any particular language are, the most one can do is to determine, using the procedure known to linguists as the comparative method, to which other languages it is related within very roughly the past four or five thousand years.[1] To discuss the question of Korean linguistic origins intelligently, one must first understand why the comparative method works and how it is applied. Let me therefore deliver a crash course that addresses both these needs.

How to establish the origins of a language

The comparative method is grounded in the observation that there are three and only three reasons that two languages show structural similarities to one another. Those reasons are chance, historical contact, and mutual inheritance of structures from an earlier common source language. Only by excluding chance and contact as plausible explanations for structural similarities can one conclude that two languages are historically related.

The scientific investigation of the world's languages, which has been conducted since the middle of the 18th century, has established not only many families of two to several dozen related languages but also super-families, or phyla, each containing two or more families. Perhaps the most thoroughly studied phylum of languages is the Indo-European, which includes almost all the languages now spoken in Europe and in large parts of Asia, including India. Many Indo-European languages had writing systems even in ancient times, so we know not only relationships among the modern members of the phylum but also a great deal about languages no longer spoken that can be traced back to the same ultimate source, called proto-Indo-European. These include the Latin of the Roman Empire, the Tocharian language spoken in what is now Xīnjiāng province in western China, and the Hittite language of ancient Anatolia, or present-day Turkey.

Written records are, however, by no means necessary for establishing relationships among languages. The great triumph of American comparative linguists of the 20th century was to establish the language families of North America, for which virtually no written records exist. The technique used to do this was essentially the same as that which had been developed in the context of Indo-European studies, but was based almost entirely on spoken data collected linguists in the field, who transcribed the speech of native speakers. The same kind of data collection and comparative analysis was also applied to the far-flung languages of the Indian and Pacific Oceans, the languages of West Asia and Africa, and the languages of China, Tibet, and Burma. Although research continues on the languages of such regions as Australia and South America, linguists already have a fairly good idea of how the languages of the world are related in terms of proto-languages that existed very roughly four or five thousands years before present. In the process of proving the existence of families and phyla through the comparative method, linguists have also identified a small number of languages, called isolates, such as Basque in

Europe and Burushaski in India, that appear to be unrelated to other known languages.

How, then, does one work with raw linguistic data to establish language relationships?

First, although it may seem that the variation in the sounds heard in everyday speech is virtually unlimited, linguists have shown that the words of every language consist of target sound categories, or phonemes, within which individual or dialectal phonetic variation does not affect the functions or meanings of the words. In any language, the identity of a word form is its unique combination of phonemes. For this reason, it possible to describe the forms of words in every language in a relatively simple and elegant way.

Second, linguists have learned that, when a phoneme of a language changes, all instances of it that occur in the same phonetic context change the same way at the same time. This is often summarized by saying that phonemic change is regular. The empirical proof of this fact by linguists in the latter part of the 19th century was a scientific breakthrough comparable in importance to the recognition of the evolution of species in biology.

Third and finally, the relationships between the phonemic structures of words and their functions or meanings in any language are entirely the product of language history. If we look at a language at a single point in time, we can equally well say that the relationships of word forms to word meanings are purely arbitrary. The only reason that, for instance, the English phoneme sequence represented by the written word *hound* (phonemically /haund/) means 'dog' is that English is a Germanic language. The Old English word *hound* meant 'dog', as does the corresponding word *Hund* (/hunt/) in High German today. And that, in turn, is because both Old English and High German descend from a pre-historic language called proto-Germanic that contained the word *hundaz meaning 'dog.' Through a series of regular sound changes, that word changed into English *hound*, German *Hund*, and all the cognate words in Dutch, Swedish, Danish, and the other Germanic languages. Likewise, if you ask why proto-Germanic *hundaz meant 'dog', it is because it arose from the pre-Germanic word *kun-tós, an enlarged form of the proto-Indo-European word *ḱwṓ from which Latin *canis* 'dog' and Greek κύων (kyōn) 'dog', and cognate words in other Indo-European languages developed.

We know these facts as a result of applying the comparative method to all the Indo-European languages we can observe or have records of. As

the name "comparative method" suggests, this involves examining words with similar forms and meanings in two languages and assessing the quality of their resemblances. Are they merely coincidental? Or are they the result of historical contact between speakers of the two languages? If both these possibilities can be excluded, then we have no choice but to conclude that the two languages are later, changed forms of a common ancestor language, perhaps no longer in existence. Even as far back as three or four thousand years before present, we are pretty sure we know when contacts between languages occurred that might have caused borrowing of words from one to another, but the primary difficulty in deciding whether two languages are related or not is excluding chance as a possible cause of resemblances. Here is where the regularity of phoneme change and the arbitrary connections between forms and meanings of words all come into play. The probability that we can find a large number of words of identical or related meanings in any two languages that also match one another phoneme by phoneme according to an exceptionless set of correspondences is extremely small. This is not only because the connections between word forms and meanings are arbitrary and because of the regularity of phonemic change but also because the joint probability of many independent events is the mathematical product of their probabilities. Even as a few as one or two hundred word matches, or etymologies, based on a compact, phonetically natural set of phoneme correspondences is enough to make it quite unlikely that the resemblances of two languages are due to chance alone.

For the sake of illustration, let's pretend that the letters of the alphabet are phonemes. The two English words *treasure* and *measure* match the Japanese words *takara* and *hakaru* in meaning; moreover, the English letter string *easur* matches the romanized Japanese string *akar* in both cases. Is this good evidence that Japanese and English are related languages? Of course not: I doubt there is even one more pair of words like these in the lexicons of English and Japanese. These matches are just a coincidence. By contrast, there are many hundreds of words in, say, Icelandic and German that have identical or similar meanings and match up, phoneme for phoneme, according to completely exceptionless rules— or as linguists often say, sound laws. The probability that such lexical matches, taken together, are coincidental is mathematically almost zero. Furthermore, given the historical record, we know that no speakers of Icelandic and German were in physical contact with one another at the time Iceland was settled by Norsemen, so we can exclude the possibility

that Old Icelandic borrowed these words from contemporary German, or vice versa. We are therefore compelled to conclude that the Icelandic and German languages are related, and that their words have been mutually inherited from a common ancestor. We would, of course, like to determine the chronological status of that common source language with respect to other languages, such as Danish, Swedish, and Dutch, that also exhibit many lexical matches with Icelandic and German, and the comparative method allows us to establish such a relative chronology. We simply apply the method systematically to each pair of languages we suspect of being related and estimate the time depth of their separation on the basis how many etymologies we find and the sound changes needed to get from the earlier word forms to the ones we observe. The result of this is a tree-like structure, the branches of which terminate in the languages we compare. The intermediate nodes where the branches bifurcate correspond to the reconstructed languages from which the observed languages arose; all the branches lead back to the root of the tree, called proto-Germanic. This reconstructed prehistoric language can, in turn, be compared with reconstructions such as proto-Romance,[2] proto-Slavic, proto-Celtic, and so on, permitting us to work our way back to a reconstruction of a proto-Indo-European language spoken millennia ago.

Of course, if two languages are related, we expect to find more than just lexical resemblances between them. We will normally also notice similarities in word structure or morphology, in sentence structure or syntax, and even in discourse structure. However, changes in these levels of language structure do not always develop with the regularity observed in phonemic change, so even two related languages may differ in morphological, syntactic, or discourse structure. For instance, French, Spanish, Italian, Portuguese, and Romanian are all later changed forms of Latin, but none retains more than faint traces of the declensions of Latin nouns and adjectives; their sentences only dimly resemble those of Latin sentences; and most of the ways people of the Roman Empire used Latin to communicate have been replaced by new habits of discourse. Furthermore, the range of variation in morphology, syntax, and discourse structure among languages of the world is much more limited than the virtually free way in which word forms are associated with word meanings and functions in any language. This is why we start with lexical comparisons rather than comparisons of higher-level structures.

How not to establish the origins of a language

Before proceeding to the question of Korean, let me emphasize that the comparative method is the only scientific procedure we have for establishing historical language relationships. Human genetics are no substitute for this methodology. We know from the study of the world's languages that groups of people with very similar DNA or physical characteristics may speak distantly related or even unrelated languages, and that people who speak the same languages may be very different physically and genetically. The ability to acquire and use language is arguably innate in all human beings, but the ability to use a particular language is the result of being reared in a particular linguistic environment or of conscious study later in life.[3]

DNA matches and differences are useful, from the perspective of linguistics, mostly when they let us exclude certain hypotheses about language relatedness. For example, we know that from about 12,000 to 2,000 years before present, Japan had a population that practiced a Neolithic culture known to archaeologists as Jōmon. Although relying primarily on fishing, hunting, and gathering of food, Jōmon communities were often prosperous; agriculture was not systematically practiced or used for subsistence, yet Jōmon communities produced the longest continuous pottery tradition of any Neolithic culture known to archaeologists. Not surprisingly, some Japanese would like to claim that the people of the long Jōmon period spoke an early form of Japanese. But it turns out that human remains from Jōmon sites are quite different from human remains found in sites of the Yayoi culture that began replacing it starting roughly 2,500 years before present. Genetically, Yayoi burials preserve bodies that are very similar to modern Japanese speakers, who, in turn, are much closer genetically to Korean speakers than either Koreans or Japanese are to other ethnic groups living in Northeast Asia. Furthermore, like people of the Mumun culture of the Korean peninsula, Yayoi folk practiced millet and wet-field rice agriculture and worked with bronze; eventually, they learned how to forge iron under influence from the peninsula. Because the introduction and rise of Yayoi culture was relatively abrupt and because of the evidence of material similarities with Mumun culture, it is widely accepted that Yayoi culture was introduced from the Korean peninsula in the first millennium BCE. Therefore, if people of the Jōmon period spoke an early form of Japanese, we are confronted with a major problem: why would a technologically more advanced group of newcomers give up their own language and adopt that

of the less advanced and far smaller Jōmon population already in the islands? The genetic and physical evidence in this case suggests we ought to reject the hypothesis that some early form of Japanese was spoken by some group of individuals in the islands during the Jōmon period.[4]

Note carefully that, while the physical and genetic evidence causes us to be skeptical of a Japanese-Jōmon linguistic connection, it by no means proves a Japanese-Korean linguistic connection. To test that hypothesis, we must apply the comparative method to linguistic data.

So too, political considerations should not be allowed to intrude upon the scientific study of language history. In South Korea, it has long been taught that the Korean language emerged exclusively from the language of the kingdom of Silla; in North Korea, the kingdom of Koguryŏ is given the honor of being the homeland of Korean. Nothing is to be gained by pursuing that dispute. From a linguistic perspective, if even though we may use the terms Sillan, Koguryŏan, and Paekchean to name the languages of the elites of the Three Kingdoms of Iron Age Korea, the degree of difference among these languages is far from clear. How many of them were related to one another? How many of them were related to Japanese? Different scholars have suggested answers to these questions ranging from zero to three,[5] and it is understandable that there should be a lack of consensus given the late date of *Samguk sagi*, which is our best source of information about the Three Kingdoms period, and the difficulties of interpreting it and the few auxiliary Chinese and Japanese sources we have available. Likewise, one should be cautious when speculating about the languages of the earlier entities called Mahan, Chinhan, and Pyŏnhan. It is often said that their languages prefigured, respectively, Paekchean, Sillan, and Wa, or pre-Old Japanese, but solid evidence of such relationships is scant.

The origins of Korean

So let us begin with modern Korean, throwing out the extensive and obvious borrowings from Chinese, and adding words we find in Middle Korean written records and those pre-*han'gŭl* materials that we can read with good confidence. We compare this body of Korean lexical material with the lexicons of each of the many languages spoken in Northeast Asia that bear rough typological similarities to Korean. The Sino-Tibetan languages like Chinese and the languages of Southeast Asia are poor candidates in this regard compared with the languages of Mongolia, Manchuria, Siberia, and, of course, Japan. In the latter part of the 19th

century, it was believed that the language families known as Uralic, Turkic, Mongolic, and Tungusic were all branches of a so-called Altaic phylum, but neither Korean nor Japanese were included as primary source (often called "witness") languages for the rigorous reconstruction of proto-Ural-Altaic. By the beginning of the 20th century, it was recognized that the number of etymologies to which Uralic languages contributed to the proto-Uralic-Altaic reconstruction was limited and not likely to expand. The odds that Uralic words that had matches in one of the Mongolic or Turkic languages were due to contact and borrowing would never significantly decrease. We do not have the same abundance of written records for Asian history that we have Europe, Northern Africa, and Southwest Asia, but we know enough of its wars, migrations, and various diffusions of religion, art, and other aspects of material culture to see that such borrowings could have occurred. The Ural-Altaic theory therefore fell into disfavor, and work on the residual Altaic theory was pursued until the latter half of the 20th century. By then, skeptical voices were suggesting that the possibility of contact and borrowing among Turkic, Mongolic, and Tungusic languages might be substantial; also, the Altaic reconstruction up to that time had neglected Korean and Japanese, which, in hindsight, seemed to be a major mistake. A similar problem had arisen in Indo-European studies when the ancient language Hittite of Asia Minor had been deciphered: one could not simply graft a Hittite branch onto the reconstructed Indo-European tree. One had instead to rework the reconstruction from the bottom up, so to speak, to determine exactly how Hittite and other ancient Asian languages split off from their linguistic relatives. So too with Korean and Japanese, it was realized that, if they were related to languages of one or more of the three Altaic families, that fact could only be proven satisfactorily by revising the entire reconstruction "from the bottom up" while putting them on an equal footing with the other languages. This is how the hypothesis of a Transeurasian phylum emerged and has come to supersede the older Altaic hypothesis. Much previous comparative work in the Altaic framework continues to have value, but its comparisons and conclusions must be re-evaluated afresh.

To give just one example, consider the kind of phonetic assimilation known as vowel harmony. Those familiar with Middle Korean know that various endings take the vowels ㅏ or ㅓ, ㅜ or ㅗ, and ㅡ or ㆍ depending on the vowels of the preceding noun or verb stem. This is an example of vowel harmony; traces of it can be found in modern Korean although they are obscured by the fact that the vowel ㆍ has merged through a series of

regular phoneme changes with other vowels in almost all dialects. Vowel harmony is found widely among Uralic, Turkic, Mongolic, and Tungusic languages. This lends some plausibility to the old Ural-Altaic hypothesis but does not come close to ruling out mere chance as a causal factor. (Languages such as Igbo, in Nigeria, also exhibit vowel harmony). Furthermore, the rules that govern vowel harmony in the Turkic languages are not the same as those at work in the Mongolic and Tungusic languages or in Middle Korean.[6] In Finnish, an Uralic language, vowels in endings assimilate in frontness to the vowels of the nouns or verbs to which they are suffixed. In Hungarian and Turkish, they also assimilate in roundedness in many cases. In Mongolic, Tungusic, and Korean, on the other hand, Seongyeon Ko demonstrates that the articulatory feature that changes due to assimilation is the position of the tongue root. Under the old Altaic theory, it had been assumed that the Turkic languages reflected the oldest stage of the common language, partly because Korean evidence was not taken into account. From the perspective of the Transeurasian hypothesis, the proto-language seems, rather, to have had tongue-root harmony, and that the change to a front/back system, sometimes involving rounding as well, was an innovation from time when the branch of Turkic languages split off from the rest of the Transeurasian family.

Returning to Korean, several scholars have attempted to compare it with Japanese and with each of the other languages of the generally accepted Tungusic and Mongolic language families. Of all these pairings, the one that has yielded the largest set of etymologies and fullest set of phonemic correspondences is Korean-Japanese. Though scholarly papers pointing out various similarities of the Korean and Japanese languages go back to the late 19th century, the first rigorous comparative results were presented by Samuel E. Martin in 1966;[7] the first major reassessment of his proto-Korean-Japanese reconstruction is found in the dissertation of John B. Whitman in 1985.[8] Alexander Francis-Ratte's 2016 dissertation is the second: he has reduced the number of regular phoneme correspondences that unify the etymologies while simultaneously increasing the phonetic naturalness of the set; eliminated etymologies that contained unexplained or doubtful correspondences, adding in even more new ones that do not; and sketched out the verb morphology of the proto-language implied by the comparison.[9] As a result, after nearly sixty years of work, we have more than five hundred rigorous etymologies that have been subjected to intensive research by many linguists showing that Korean and Japanese sprang from a common ancestor language. Questions

have been raised about this or that etymology, which is only to be expected in a scientific enterprise, but broad objections, such as the claim that all the resemblances between Korean and Japanese are due to historical contact,[10] are no longer credible. This is because only three or four dozen words can be securely identified as borrowings from Old Korean into Old Japanese in the 7th and 8th centuries, when refugees came to the islands of Japan from the Korean peninsula following the conquests of Silla. In short, if we focus strictly on the proven scientific method for determining language relationships, the only affirmative conclusion we have so far is that there once existed a proto-Korean-Japanese language from which Korean and Japanese both diverged in prehistoric times.

The forms shown in Tables 1 through 3 give an idea of the pKJ reconstruction. It will be observed that the meanings of the nouns or verb roots compared are not always identical, but that, when they differ, there is a plausible explanation for the change of meaning or function from the one found in the proto-language to the one reflected in the daughter branch.

Table 1: Some pKJ Etymologies

Old Japanese	proto-Japanese	Middle Korean	proto-Korean	proto-Korean-Japanese
yama 'mountain'	*jama	*yém* 'rocky island'[11]	*jəm / *jem	*jəma 'island'
na-/ne 'root'	*naj	*nol* 'raw'	*nər	*nar 'root'
po-/pwi 'fire'	*pəj	*púl* 'fire'	*pɪr	*pir 'fire'
wata 'ocean'	*wata	*patáh/palól* 'ocean'	*pata + *kə '(locative)'	*wat-a 'ocean'
ko- 'comes'	*kə-	*ká-* 'goes'	*ka-	*kə- 'comes'
wor- 'is at'	*wo- 'comes'	*wó-* 'comes'	*wo-	*wo- 'comes'
nar- 'becomes'	*na- 'goes out'	*ná-* 'goes out'	*na-	*na- 'goes out'

Source: Ratte 2016.

On the other hand, the phonemic correspondences necessary for the matching of the compared forms all belong to the small, phonetically natural, and regular sets shown in Tables 2 and 3.

Table 2: Principal Consonant Correspondences

proto-Korean-Japanese	proto-Japanese	proto-Korean
*p	*p	*p
*t	*t	*t
*c	*t / *s	*c
*k	*k	*k
*s	*s	*s
*x	*k / *s	*h
*m	*m	*m
*n	*n	*n
*ŋ	*ŋk = *ŋ	*ŋ
*r	*r	*r
*rr	*j	*-rr-/-rər-
*j	*j	*j
*w	*w	*p

Source: Ratte 2016.

Table 3: Principal Vowel Correspondences

proto-Korean-Japanese	proto-Japanese	proto-Korean
*a	*a	*a
*ɨ	*ə	*ɨ
*ə	*ə	*ə
*o	*o	*o
*u	*u	*u
*i	*i	*i
*e	*ə / ___[+COR], *e	*e / ___[+COR], *je

Source: Ratte 2016.

By contrast, the sampling of words shown in Table 4, which are clearly borrowed from Korean into Japanese, almost all in first millennium CE, do not adhere strictly to these correspondences. It is also obvious that they are words referring to aspects of culture unknown in either pre-Korean or

pre-Japanese times, or that their meanings in Japanese are much narrower than their meanings in Korean.

Table 4: Words Borrowed from Korean into Japanese

OJ *kusiro* 'type of bracelet' ~ MK *kwusúl* 'jewel'
OJ *kisi* (title for Korean lord) ~ MK *kuwí(síl)* 'government post,' OK (Pc) *kisi* 'lord'
OJ *kisaragi* '如月 2nd month' ~ MK *kyezúlh* 'winter'
OJ *kimi* 'lord' ~ MK (*ni:m*)-*kúm* 'lord'
EMJ (*asa*)-*borake* 'dawn' ~ MK *polk/pulk*- 'bright, red'
OJ *yorokob*- 'rejoices' ~ MK *cúlkeW*- 'joyful'
OJ *potoke* 'Buddha' ~ MK *pwuthye* 'Buddha'
EMJ *saki-kusa* '三枝'10 ~ MK *se:yh* '3' + *káci* 'branch'
OJ *kama* 'pot' ~ MK *káma* 'pot'
OJ *sasi* 'fortress (lexicalized)' ~ MK *cás* 'fortress,' OK (Pc) *sasi* 'fortress'
OJ *iraka* 'roof tile' ~ MK *iráng* 'furrow'
OJ *karamusi* 'Chinese ramie' ~ MK *mwosi* 'ramie cloth'
OJ *kasa* 'straw umbrella' ~ MK *kás* 'hat, covering'
EMJ *sarapi* 'rake' ~ MK *sálp* 'spade'
EMJ *sitogi* 'ceremonial rice cake' ~ MK *sték* 'rice cake'
OJ *kopori* 'district' ~ MK *kwowolh* 'district', OK (Pc) *kopori* 'district'
OJ *tera* 'temple' ~ MK *tyél* 'temple'
OJ *pyera* 'moldboard, spatula' ~ MK *pyés* '(cock's) comb'
OJ *pati* 'bowl' ~ MK *palí* 'bowl' (from Sanskrit?)
OJ *kura* 'valley (in toponyms)' ~ MK *kwo:l* 'valley'
OJ *kure* 'Korea, Koguryŏ' ~ K *kwokwulye* 'Koguryŏ'
EMJ *kure* 'measured piece of wood' ~ MK *kuluh* < PK *kul* 'wood'
OJ *suki* 'baby-carrying sling' ~ MK *kính / skinh* 'cord'
OJ *sa*- / *sati* / *satu* 'arrow' ~ MK *sál* 'arrow'
J *kugutu* 'puppet' ~ K *kkwoktwu* 'puppet'
OJ *kinu* 'silk' ~ K *kyen* 'silk' (Chinese *juàn* 絹)
EMJ *sozoro* 'involuntarily' ~ MK *susúy-lwo* 'on one's own'

OJ *suwe* 'Korean pottery' ~ MK *swolá* 'dish, vessel'	
OJ *mure* 'mountain (lexicalized)' ~ MK *mwo:yh* 'mountain,' OK (Pc) *mure* 'mountain'	

Source: Ratte 2016.

In these loanwords, we find the following irregular correspondences.

Table 5: Consonant Correspondence in Loanwords

Old Japanese	Middle Korean	Proto-Korean
z	*s / z*	*s
gV	*k#*	*kV
rVgV	*lh*	*rVk
rVpV	*lp*	*rVp
rVkVb	*lkVW*	*rVkVp
rVkV	*lk*	*rVk
r	*t*	*t
we	*la*	*ra
Ø	*-n / -nh*	*n

Source: Ratte 2016.

Table 6: Vowel Correspondence in Loanwords

Old Japanese	Middle Korean	Proto-Korean
(C)wo (non-final)	*(C)wo*	*o
(C)ye (non-final)	*(C)ye*	*e / *je
e (non-final)	*ye*	*e
a#	*a#*	*aCa
a	*u*	*ɨ
i	*u*	*ɨ
i	*wi / uwi*	* ɨwi / *uwi
u	*u*	*ɨ

Source: Ratte 2016.

In short, from the perspective of linguistics, the most secure deduction we can draw at present about the origins of the Korean language is that it shares a common source with the Japanese language. This naturally raises

a number of questions, some linguistic and some belonging to the realm of history, archaeology, human genetics, and other academic disciplines. Studies such as Seongyeon Ko 2018, discussed earlier, help us limit the range of hypotheses we wish to entertain about the relationship of proto-Korean-Japanese with possibly related languages of the Mongolic and Tungusic families, but, because they deal with purely linguistic facts, they are not of much help in deciding where or when this or that speech community existed. In what follows, I will summarize my current understanding of the location and history of proto-Korean-Japanese and how it may be related to other prehistoric languages of Northeast Asia, which is broadly informed by the archaeological work of Miyamoto Kazuo (2016, 2019) and Whitman 2011.[12]

Situating proto-Korean-Japanese

A language we may call Macro-Tungusic was probably spoken by people dwelling on the western coast of the Bay of Bóhǎi (modern Tiānjīn municipality and the adjacent coastal areas of Héběi province) roughly 5,000 years before present. In the third millennium BCE, some members of this group migrated northeastward toward Lake Khanka in what is now the Russian Maritime; their language become proto-Tungusic, from which the Tungusic languages of East Siberia diversified. (If one accepts the Transeurasian hypothesis, Macro-Tungusic was a first-order sister of the proto-language from which the Turkic and Mongolic languages later emerged).

The migration of pre-Tungusic speakers left behind a group whose language we can properly call proto-Korean-Japanese; because communications between the northern and southern parts of this group were increasingly attenuated by physical separation, their dialects diverged into pre-Korean and pre-Japanese varieties. In the second millennium BCE, people from the northern, pre-Korean subgroup moved eastward into the Yalu River valley, later spreading south along the western coast of the Korean peninsula. People of the southern or pre-Japanese subgroup moved along the southern side of the Bay of Bóhǎi reaching the Shāndōng peninsula, from which they crossed by sea to the Liáodōng peninsula. From there, they too spread into the Korean peninsula. The influx of these migrants and the technologies they brought with them (millet agriculture, rice farming, bronze, and dolmen burials) signaled the transition from Chǔlmun to Mumun culture.

During the Mumun period, a cooling climate event stimulated all the peoples inhabiting the peninsula to venture southward. This prompted many pre-Japanese speakers in the south to establish permanent settlements in northeastern Kyūshū and southwestern Honshū. Miyamoto finds archaeological evidence that this happened in at least two phases: first, between the 9th and 8th centuries BCE; and then between the 7th and 6th. Although rice agriculture was introduced during Phase 1, Miyamoto argues that the beginning of Yayoi culture in Japan should be associated with the evidence found at the archaeological site of Itazuke dating from the 6th to 5th centuries, which shows features of the mature form of the Phase 2 culture.

With the emigration of many pre-Japanese speakers from the Korean peninsula, pre-Korean speakers became increasingly dominant there. They were the people most directly influenced by contact with the expanding Hàn empire of China. Koguryŏ was the first sinified kingdom whose rulers probably spoke Old Korean; the linguistic status of the Puyŏ people to their north is less clear. At any rate, the southward drift of Koguryŏ culture took the form of the vanguard kingdom of Paekche in the southwestern peninsula. The establishment of Chinese commanderies temporarily interrupted communication between Koguryŏ and Paekche and afforded it political autonomy; after the commanderies fell to Koguryŏ, a second vanguard kingdom, Silla, emerged in the southeast.[13]

It is important to keep in mind that the pre-Japanese speakers who took their languages to the islands were not organized into cohesive states, though the names Mahan, Chinhan, and Pyŏnhan recorded in Chinese histories about the general area they inhabited may make it seem otherwise. Even as late as the formation of Silla, archaeological evidence shows social complexity in the Yŏngnam region,[14] possibly reflecting the existence of Old Korean-speaking groups next to speakers of para-Japanese, as we call the peninsular language left behind by the Yayoi migrants. Certainly migrants from the peninsula who settled in the area of Izumo were aware of their connections with the Yŏngnam region, of which they left traces in many of the Japanese myths about Susa-no-wo-no-mikoto and his progeny Opo-kuni-nusi. In the case of Paekche, relations between the rulers, who spoke Old Korean, and those of their subjects who spoke para-Japanese—Chinese sources tell us two languages were spoken in the kingdom—seem to lie at the root of Paekche's unusually close relationship with the emerging kingdom of Yamato.

Various researchers have assumed that, because there were three politically distinct kingdoms—Koguryŏ, Paekche, and Silla—each had its own language, but, for a variety of reasons (reviewed in Unger 2014), I think this is mistaken.[15] The simplest hypothesis—that all the rulers spoke varieties of Old Korean—is the most likely not only because of Occam's Razor but also because, after the military destruction of Koguryŏ and Paekche in the 7th century, multilingualism seems not to have been one of the problems facing the kingdom of Unified Silla. On the contrary, though primary documentary evidence from this period is extremely limited, it seems that a prince of Paekche had no trouble composing a *hyang-ka* in Old Korean.

Let me emphasize that this description is tentative. As far as linguistic data are concerned, we know that Korean and Japanese are related and that the languages of the rulers of Koguryŏ, Paekche, and Silla were related, though noticeably different by the time they appear in history as Chinese-style kingdoms. There is some debate as to whether proto-Korean-Japanese is better understood as a member of a Macro-Tungusic family, as I have described it, or as a branch of a Transeurasian phylum coordinate with proto-Turkic, proto-Mongolic, and proto-Tungusic. Future research will tell us which branching model is superior. At this stage, the foregoing summary is the closest we can come to describing the origins of the Korean language.

Notes:

[1] The limit on time depth is a consequence of the fact that all languages change over time, and it has been empirically observed that virtually no distinctive traces of a language survive more than about five thousand years.

[2] Proto-Romance was, more or less, the kind of Latin spoken in the streets of the late Roman Empire, and should not be confused with the classical Latin, still taught in schools today, of many centuries earlier.

[3] The linguistic environment can be quite complex. In many traditional societies, bilingualism is common. There are even societies that practice linguistic exogamy together with patrilocal marriage: a man must marry a woman whose childhood language is different from his, and bring her to his home.

[4] J. Marshall Unger, "No rush to judgment: the case against Japanese as an isolate," *NINJAL Project Review,* Volume 4, Number 3 (2014), pp. 211-230.

[5] Christopher I. Beckwith, *Koguryo, the language of Japan's continental relatives: An introduction to the historical-comparative study of the Japanese-Koguryoic languages with a preliminary description of Archaic northeastern Middle Chinese* (Leiden, Netherlands: Brill, 2007); J. Marshall Unger, *The Role of Contact in the Origins of the*

Japanese and Korean Languages (Honolulu: University of Hawai'i Press, 2009); and Ki-moon Lee and S. Robert Ramsey, *A History of the Korean Language* (Cambridge: Cambridge University Press, 2011).

[6] Soengyeon Ko, *Tongue root harmony and vowel contrast in Northeast Asian languages*, Turcologica 112 (Wiesbaden, Germany: Harrassowitz, 2018).

[7] Samuel E. Martin, "Lexical Evidence Relating Korean to Japanese," *Language*, Volume 42, Number 2 (1966), pp. 185-251.

[8] John B. Whitman, The Phonological Basis for the Comparison of Japanese and Korean, Ph.D. dissertation, Harvard University, 1985.

[9] Alexander T. Francis-Ratte, Proto-Korean-Japanese: A New Reconstruction of the Common Origin of the Japanese and Korean Languages, Ph.D. dissertation, The Ohio State University, 2016.

[10] Alexander Vovin, *Koreo-Japonica: A Re-evaluation of a Common Genetic Origin* (Honolulu: University of Hawai'i Press, 2010).

[11] Cf. MK *yém-sywó* 'goat,' *yém-kywo* 'scallon,' NK *yem* 'small rocks sticking out of the water' < pre-MK *yém 'mountain'. This rejects the etymology of *yém-sywó* as 'bearded cow' as a folk etymology; *(swu)yem* is a borrowing from Chinese.

[12] Kazuo Miyamoto, "Archaeological Explanation for the Diffusion Theory of the Japonic and Koreanic Languages," *Japanese Journal of Archaeology*, Volume 4 (2016), pp. 53–75, and "The Spread of Rice Agriculture during the Yayoi Period: From the Shandong Peninsula to the Japanese Archipelago via the Korean Peninsula," *Japanese Journal of Archaeology*, Volume 6 (2019), pp. 109–124; John B. Whitman, "Northeast Asian Linguistic Ecology and the Advent of Rice Agriculture in Korea and Japan," *Rice*, Volume 4 (2011), pp. 149–58.

[13] John B. Whitman, "Northeast Asian Linguistic Ecology and the Advent of Rice Agriculture in Korea and Japan."

[14] Jack Davey, "Unreliable narratives: historical and archaeological approaches to early Silla," *Seoul Journal of Korean Studies*, Volume 29, Number 1 (2016), pp. 7–32 and Jack Davey, "Culture contact and cultural boundaries in Iron Age Southern Korea," *Asian Perspectives*, Volume 58, Number 1 (2019), pp. 123–48.

[15] J. Marshall Unger, "No rush to judgment: the case against Japanese as an isolate**."**