https://doi.org/10.3176/lu.1994.3.01

REIN TAAGEPERA (Irvine—Tartu)

THE LINGUISTIC DISTANCES BETWEEN URALIC LANGUAGES*

Fair agreement exists among linguists that Samoyedic and Finno-Ugric split around 4000 BC at the latest, but later branching dates are controversial. The split of Ugric and Finno-Permic is set around 3000 BC by Korhonen (1981) but 2000 BC by Hajdú (1975 : 42 and 1976 : 39). The split of Volgaic and Finno-Samic is seen around 1500 BC by Korhonen but between 1000 and 500 BC by Hajdú. And so on. This article has two purposes.

1) Based on the degree of similarity of basic vocabulary reported by Alo Raun (Raun et al. 1965), I estimate the separation dates of various languages, using a simple exponential attrition model. The most striking result is an almost simultaneous parting of ways of Permic, Mari, Moksherzian, and Finno-Samic groups between 1500 and 1150 BC.

2) Based on the same data, the article further develops a graphical method to visualize the distances between languages. The closeness of even the most distant Turkic languages offers a visual contrast to the large distance between such relatively close Uralic languages as Mari and Moksherzian. The validity of the conclusions depends of course on the adequacy of the starting data.

It should be noted that all of the preceding is amenable to refinement and extension of data. The exponential method in part 1 can be extended to other cases (such as the split between southern and northern Samoyedic or that between Samic and Finnic), once standard word lists are established for further languages (such as Selkup and Samic). The same applies to the graphical method in part 2. The limitations of these approaches are discussed in the course of the article, but the broad nature of preassumptions involved should be indicated right away.

Part 1 implies a tree model of language affinity that oversimplifies and possibly distorts a vastly more complex process of linguistic interactions (cf. Sinor 1988 : xiii—xx). This is what simple models do. They merely say that, within bounds to be specified, nature behaves ”as if” the model applied. When faced with inconsistent cases, the simple model would have to be either made more complex in a careful way or discarded in favor of another, more efficient simple model. Giving up on a model without replacing it and merely declaring that things are complex and fuzzy does not advance science. This is why I continue here to refine the tree model, offering a more systematic method to determine the language separation dates, while fully aware that separation dates at best represent an average date during a long and fuzzy period, and at worst reflect no

* I thank Ago Künnap, Alo Raun and Ott Kurs for comments and advice.

historical reality whatsoever apart from "as if." Part 2 is completely independent of the language family tree model, or any other model of the past. It only helps us visualize the present affinities of vocabularies involved, regardless of how the particular languages came about (and how the set of words chosen reflects the languages).

Data Base

Table 1 shows the percentage of basic words that are common to various Uralic languages, as reported by Alo Raun, who made use of the standard comparison list of 100 words proposed by Swadesh (Raun et al. 1965 : 33). This is the data set employed here. Presumably Meadow Mari and Erzian are used, although the source does not specify it. I have added columns for average commonalities of Nenets and Hungarian with the Finno-Permic languages, and for that of all other languages with the Volgaic languages. Finally, the last column represents the average commonality of the given language with all the others. Thus Nenets has, on the average, 14.6% words in common with the other Uralic languages listed.

Raun's table dates 30 years back, and further critical work on Uralic vocabulary has been published since (Janhunen 1981, in particular). However, no update of a comprehensive table like Raun's seems to have been published. A future revision is likely to change some percentages in Table 1 in an unpredictable direction and to an unpredictable degree. For the moment, Raun's data are the only basis available.

According to Table 1, the average commonality of Finno-Permic languages with Hungarian (27%) is not much lower than the commonalities of Mari and Moksherzian (36%), although the latter traditionally are placed in the same Volgaic group. Only with the Samoyedic languages do we go a marked notch further down in commonality of words (11 to 19%, depending on the particular Finno-Ugric language and including some random variation).

Further comparable data are given by Lehtiranta (1982, as cited by Sammallahti 1988 : 499): in the Swadesh 100-word list, Hanti-Mansi commonality is 45, while for Mansi-Hungarian it is 34 and for Hanti-Hungarian 28. Thus the Hungarian-Obugric commonality is 31+3%, which is barely higher than the commonality of Hungarian with the Finno-Permic languages.

The general averages shown in the last column suggest that Mari tends to be the central Uralic language in the sense that it has most in common with the other languages. It may indicate that Mari is the most conservative among the Uralic languages, or it might be an artifact of the particular list of words and choice of languages in Table 1.

Table 1

Percentage of basic words common to Uralic languages
(Based on Raun et al. (1965), using the 100 standard words proposed by Swadesh)

						Finno-Permic	Volgaic	General
	Hung.	Komi	Finnish	Mari	Erzian	Average	Average	Average
Nenets	13	11	15	19	15	15.0	17.0	14.6
Hungarian		26	27	30	25	27.0	27.5	24.2
Komi			31	40	27	—	33.5	27.0
Finnish				36	34	—	35.0	28.6
Mari					36	—	—	32.2
Erzian						—	—	27.4

Separation Dates Suggested by Exponential Attrition Model

The percentage of common words in Table 1 can be used to estimate the time durations the various language groups have been separate. If one assumes that renewal of vocabulary proceeds at an approximately steady pace, then C, the percentage of common words, would decrease exponentially over time:

$$C = 100 \exp(-t/T),$$

where t is time in years and T is a characteristic time interval during which C is reduced by a factor of 1/e=.37 (e being the basis of natural logarithms). There are of course periods during which certain languages have undergone extensive change, then remained relatively quiescent. Over a long period, some languages undergo less change than others. (Mari would seem relatively conservative, on the basis of Table 1.) We'll address this issue later on. As a first approximation, it will be assumed that, on the average over time, all Uralic languages have lost common vocabulary at the same rate, corresponding to the same characteristic time T. This characteristic time is determined as follows.

There is some consensus that the split of Finno-Ugric from Samoyedic occurred 6000 years ago at the latest. If we plug t=6000 and C=14.6 (average commonality of Nenets with Finno-Ugric languages, in Table 1) into the equation above, the result is T=3120 years. The general equation becomes

$$C = 100 \exp(-t/3120),$$

which can be rearranged as

$$t = -3120 \ln(C/100).$$

The latter equation can be used to calculate the separation time for any two languages, when their percentage of common vocabulary (C) is given. However, keep in mind the anchor point of 6000 years for the split of Samoyedic and Finno-Ugric. If that split is placed earlier (say, 7000 or 8000 years ago), all other separation times would become proportionately longer too.

Table 2

Time distances between Uralic languages

	Common[a] vocabulary (%)	Time[b] Distance (years)	Separation Date (B.C.)		
			Calculated[c]	Hajdú 1975; 1976	Korhonen 1981
Samoyedic **from Finno-Ugric**	14.6	6,000	4,000	4,000	4,000
Ugric **from Finno-Permic**	27.0	4,100	2,100	2,000	3,000
Permic **from Finno-Volgaic**	32.2	3,500	1,500	1,500	2,000
Finnic **from Volgaic**	35.0	3,300	1,300	1,000—500	1,500
Mari **from Moksherzian**	36	3,200	1,200	100	1,500
Hungarian **from Ob-Ugric**	31[d]	3,600	1,600	500	1,000
Mansi **from Hanti**	45[d]	2,500	500	—	—

[a] From Table 1.

[b] Calculated from *a*, using Eq. 2, except for Samoyedic from Finno-Ugric, which is preassumed.

[c] From *b*.

[d] From Lehtiranta (1982).

The results are given in the top part of Table 2, which lists the values of C for various languages (or averages for language groups) and the resulting distances in time, rounded off to closest full century. The next column gives the corresponding separation dates. For comparison, separation dates proposed by Hajdú (1975 : 42 and 1976 : 39) and Korhonen (1981) are also given. For the Ugric-Finnic and Permic-Finnic splits, the exponential approximation comes closer to Hajdú's estimates. However, for the Volgaic-Finnic and Moksherzian-Mari split it comes closer to Korhonen (Korhonen does not give explicitly the latter date, but his sketch suggests that a common Volgaic phase did not exist).

The random error range on our results can be estimated by applying the model to Hungarian and the individual Finno-Permic languages. While their average C yields a separation date of 2100 BC, Erzian alone would give 2300, Komi 2200, Finnish 2100, and Mari 1750 BC. Hence the random error range is plus or minus 200 years. Within this range, Finnic, Mari, Moksherzian, and Permic could all have split off from each other simultaneously; at the other extreme, the Volgaic common phase could have lasted for up to 300 years — but hardly more.

As shown at the bottom of Table 2, Lehtiranta's (1982) Ugric figures lead to a Hungarian-Obugric split around 1650 BC, plus or minus 300 years. This is much earlier than either Hajdú's or Korhonen's estimates but agrees with Sammallahti's (1988). Indeed, within the range of error, the Obugric languages could have split away from Hungarian simultaneouly with the Finno-Permic. (I do not suggest that this was the case.) The date of Hanti-Mansi split around 500 BC tends to agree with Sammallahti, who suggests 1000 BC, in contrast to some others who place it as late as AD 1200 (Uibopuu 1984 : 261).

In sum, the exponential approximation suggests that common Finno-Ugric lasted for 2000 years, but then split into six separate groups within less than one millennium (2100 to 1200 BC) — see graphical sketch in Figure 1: Hungarian, Obugric, Permic, Mari, Moksherzian, and Finno-Samic (which itself split soon after). All this depends on the representativeness of Raun's and Lehtiranta's samples — and on the uniformity of the attrition process. The latter issue will now be discussed.

The simple model is formally set up as if the languages were joined one year and lost all contact the next. This of course hardly took place, short of a sudden move of a thousand kilometers. The usual pattern was a gradual loosening of geographical contact or of linguistic interaction, due to decreasing intelligibility of diverse dialects. In some cases (such as southern and northern Estonian or western and and eastern components of Finnish) renewed fusion of related languages took place. Even in the absence of refusion, common Uralic word roots lost in a language could be reinserted through later loans from another Uralic language. The exponential model does not deny such processes but simply yields an average date for a slow separation (or separation-refusion) period.

After separation, language change can proceed unevenly. Take a rather extreme case, a hypothetical language somewhat like Hungarian. Assume that a mixture of Ugric and Turkic populations leads within 200 years to the infusion of one-fifth Turkic words into the basic vocabulary. According to the exponential equation above, such a decrease to C=80% of the previous Ugric vocabulary is expected to take place over 700 years. Hence we would overestimate the separation time of Ugric from Finno-Permic by some 700–200=500 years. If the above case applied to the real Hungarian, it could be that Ugric did not separate from Finno-Permic any earlier than Permic but simply underwent a period of rapid change later on (I do not suggest this was the case).

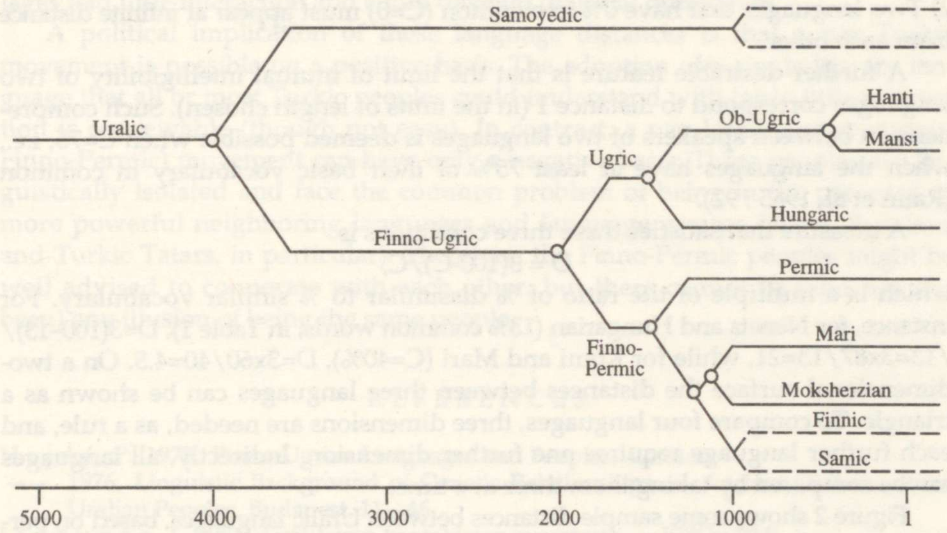

Figure 1. **Differentiation of Uralic languages over time B. C.** (based on Table 2)

In other words, an unusually heavy impact by a foreign population would be translated into temporal terms. The question is: what amounts to "unusual"? Over the course of millennia, all Finno-Ugric languages have undergone interactions with neighbors plus internal changes to a fairly equal degree , as shown by the degree of commonality of vocabulary (24 to 32%, according to the last column in Table 1). This average impact has already been taken into account by the characteristic time T=3120 years. Therefore, the basic usefulness of the exponential model is preserved, but the results should not be taken for granted with a precision of one or even three centuries. The results of this method must be taken in conjunction with those of other approaches.

It would be of interest to extend Raun's table to other Uralic languages. For one, exponential estimates could then be given for separations such as those between southern and northern Samoyedic, Samic and Finnic, Komi and Udmurt, and Mokshan and Erzian. Such work would also establish firmer limits of credibility for the exponential method, in the following way. If Obugric vocabulary should lead to the same Ugric-Finnic separation date as does the Hungarian, then the credibility of this date obviously would be boosted. On the other hand, if disagreement should surpass 200 or 300 years (the random error range established earlier), other explanations would have to be considered, such as a possibly faster attrition in some languages. However, such extension of Raun's work is beyond the present author's capability.

Visual Representation of Language Differences

It would be desirable to have a way to show graphically how far two or three related languages are from each other. I propose here such a method and will use it with various Uralic and also Turkic languages.

Any such visualization must satisfy two "boundary conditions":

1) Two languages that have 100% in common (C=100, in the preceding notation) must appear at zero distance from each other.

2) Two languages that have 0% in common (C=0) must appear at infinite distance from each other.

A further desirable feature is that the limit of mutual intelligibility of two languages correspond to distance 1 (in the units of length chosen). Such comprehension between speakers of two languages is deemed possible when C=75, i.e., when the languages have at least 75% of their basic vocabulary in common (Raun et al. 1965 : 92).

A measure that satisfies these three conditions is

$$D = 3(100\text{-}C)/C,$$

which is a multiple of the ratio of % dissimilar to % similar vocabulary. For instance, for Nenets and Hungarian (13% common words, in Table 1), D=3(100–13)//13=3x87/13=21, while for Komi and Mari (C=40%), D=3x60/40=4.5. On a two-dimensional surface the distances between three languages can be shown as a triangle. To compare four languages, three dimensions are needed, as a rule, and each further language requires one further dimension. Indirectly, all languages can be compared by taking them three at a time.

Figure 2 shows some sample distances between Uralic languages, based on percentages in Table 1. Finnish, Erzian, Mari and Komi appear practically equidistant. Hungarian appears barely more distant from Finno-Permic languages than these are from each other, or Hungarian from Obugric. In contrast, Nenets Samoyedic is very distant from Finno-Ugric languages. (It cannot be concluded from Figure 2 that Nenets is somewhat closer to Finno-Permic than to Ugric; this is likely to be random variation, which is magnified when D is large.) Circles have been drawn in at the distance of mutual intelligibility (C=75; D=1). The visual representation drives in the lack of such intelligibility between Finno-Ugric subgroups shown.

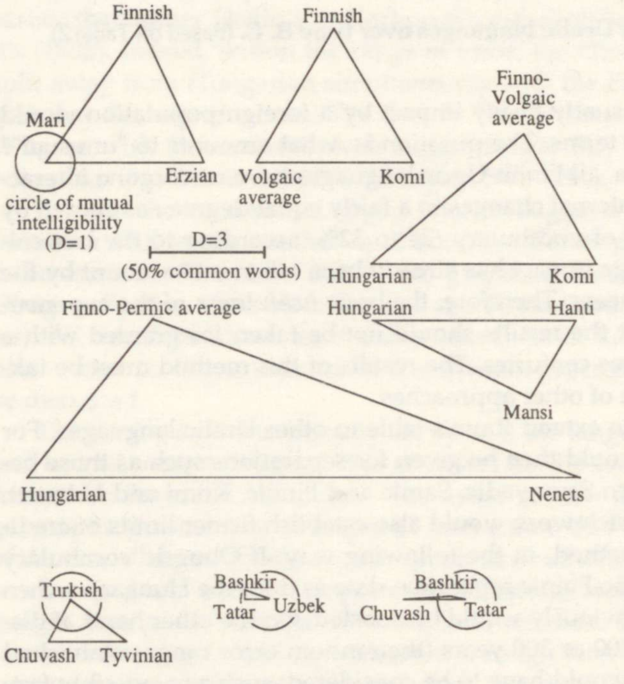

Figure 2. **Distances of Uralic and Turkic languages** (D=3% dissim./% similar)

For comparison, some relationships between Turkic languages are also shown on the same scale, based on data presented by Raun et al. (1965 : 92). Mutual intelligibility is the case among several Turkic languages within the former USSR (Tatar, Bashkir and, marginally, Uzbek); even the relatively isolated Chuvash and Tyvinian have 60% or more of their vocabulary in common with Tatar (and with the Turkish proper, in Turkey). There is a dramatic visual difference between the semi-intelligibility of most Turkic languages and the lack of intelligibili-ty among the Uralic subgroups. The Turkic people used to roam the steppes on horseback, resulting in frequent contacts and remixing. The Uralic people tended to be localized hunters and farmers, a way of life where even nearby villages had little interaction and could develop dialectal differences.

A political implication of these language distances is that a pan-Turkic movement is possible on a positive basis: The adoption of a single literary language that all or most Turkic peoples could understand with fairly little instruction is conceivable (though not easy). In contrast, a pan-Uralic (or even pan-Finno-Permic) movement can have only a negative basis: These peoples are linguistically isolated and face the common problem of being in the presence of more powerful neighboring languages and language groups (Slavic Russians and Turkic Tatars, in particular). Therefore, the Finno-Permic peoples might be well advised to cooperate with each other, but there cannot be (and has not been) any illusion of being the same people.

REFERENCES

Hajdú, P. 1975, Finno-Ugrian Languages and Peoples, London.

—— 1976, Linguistic Background of Genetic Relationships. — Ancient Cultures of the Uralian Peoples, Budapest, 11—46.

Janhunen, J. 1981, Uralilaisen kantakielen sanastosta. — JSFOu 72 9.

Korhonen, M. 1981, Johdatus lapin kielen historiaan, Helsinki.

Lehtiranta, J. 1982, Eine Beobachtung über die Gründe der raschen Veränderung des Grundwortschatzes im Lappischen. — FUF 44, 114—118.

Raun, A., Francis, D., Voegelin, C. F., Voegelin, F. M. 1965, Languages of the World: Boreo-Oriental, Fascicle One. — Anthropological Linguistics 7 1.

Sammallahti, P. 1988, Historical Phonology of the Uralic Languages. — The Uralic Languages: Description, History and Foreign Influences, Leiden, 478—535.

Sinor, D. (ed.) 1988, The Uralic Languages: Description, History and Foreign Influences, Leiden.

Uibopuu, V. 1984, Meie ja meie hõimud, Lund.

РЕЙН ТААГЕПЕРА (Ирвин—Тарту)

ЛИНГВИСТИЧЕСКИЕ ДИСТАНЦИИ МЕЖДУ УРАЛЬСКИМИ ЯЗЫКАМИ

Если предположить, что общность самодийских и финно-угорских языков распалась в IV тыс. до н. э. и что позднее рассеяние их общей лексики следовало экспоненциальной модели, то степень совпадения 100 слов по Сводешу дает следующую картину: выделение всех шести основных финно-угорских ветвей (обско-угорская, венгерская, пермская, марийская, мордовская, прибалтийско-финско-саамская) происходило в течение менее чем тысячелетия, в 2100—1200 гг. до н. э. (табл. 2 и рис. 1). Особенно неожиданным оказалось то, что разделение между собой обско-угорских и венгерского языков свершилось, вероятно, даже раньше (в 1600-е годы до н. э.), чем пермских и волжско-финских. Во второй части статьи описан метод, который позволяет наглядно продемонстрировать расстояния между тремя языками. Некоторые примеры (рис. 2) четко показывают, как далеки между собой даже близкие уральские языки в сравнении с тоже далекими между собой такими тюркскими языками, как турецкий и тувинский.