A Central Asian language survey: Collecting data, measuring relatedness and detecting loans

Philippe Mennecier (1), John Nerbonne (1,2), Evelyne Heyer (1) and Franz Manni (1,2)[1]

(1) Musée de l'Homme, Department 'Hommes, Natures, Societés', Eco-Anthropologie et Ethnobiologie, UMR 7206 CNRS, Univ Paris Diderot, Sorbonne Paris Cité, F-75005, National Museum of Natural History, Paris, France.
(2) Department of Humanities Computing, University of Groningen, Groningen, The Netherlands.

Abstract:

We have documented language varieties (either Turkic or Indo-European) spoken in 23 test sites by 88 informants belonging to the major ethnic groups of Kyrgyzstan, Tajikistan and Uzbekistan (Karakalpaks, Kazakhs, Kyrgyz, Tajiks, Uzbeks, Yagnobis). The recorded linguistic material concerns 176 words of the extended Swadesh list and will be made publically available with the publication of this paper.

Phonological diversity is measured by the Levenshtein distance and displayed as a consensus bootstrap tree and as multidimensional scaling plots. Linguistic contact is measured as the number of borrowings, from one linguistic family into the other, according to a precision/recall analysis further validated by expert judgment.

Concerning Turkic languages, the results of our sample do not support regarding Kazakh and Karakalpak as distinct languages and indicate the existence of several distinct Karakalpak varieties. Kyrgyz and Uzbek, on the other hand, appear quite homogeneous. Among the Indo-Iranian languages, the distinction between Tajik and Yagnobi varieties is very clear-cut.

More generally, the degree of borrowing is higher than average where language families are in contact in one of the many sorts of situations characterizing Central Asia: frequent bilingualism, shifting political boundaries, ethnic groups living outside the "mother" country.

1. Introduction

The primary purpose of this paper is to survey and document the linguistic relations among a genealogically mixed group of languages in Central Asia, some Turkic and others Indo-Iranian. This is a descriptive goal. Methodologically we have the modest aim of using a measure of pronunciation (dis)similarity as an inverse measure of relatedness within the two genealogical families. We measure pronunciation dissimilarity using Levenshtein distance, which has seen a great deal of use in dialectology (Wieling and Nerbonne 2015), but less in assaying relations at a great historical depth. But using the Levenshtein distance in this way is definitely not innovative. The Automated Similarity

[1] Correspondence: Dr. Franz Manni, UMR 7206, National Museum of Natural History - Musée de l'Homme, CP 139, 57 rue Cuvier, 75231 Paris Cedex 05, France Email: franz.manni@mnhn.fr

Judgment Program (AJSP) demonstrates the usefulness of the Levenshtein distance (also known as edit distance) in historical inference (Wichmann et al. 2013), however (see also Jäger 2013). Se we expect to be successful in detecting relations within the two language families. We note that the effort is also interesting because of the close contact among the peoples of central Asia; this complicates the situation. We likewise incidentally report that the pronunciation distances derived from the 200-word Swadesh list are essentially indistinguishable from those derived from the 100-word list.

We also document borrowing among these languages as a likely reflection of contact (including indirect contact) and we try out the obvious idea of recognizing loan words by unexpectedly similar pronunciation as measured by Levenshtein distance. We emphasize here and below that while we can detect loan words, out methods cannot distinguish direct loans from loans that enter the language via a third variety. But it turns out that we can indeed recognize loan words, albeit imperfectly.

In the subsection immediately following we sketch the background of the project, which was initiated by population geneticists. We note this here because of the effect it had on the sample of language varieties studied. This is followed by sections on methodology and on results, and we conclude with a discussion section.

1.1. Background

Since thirteen years ago and within diverse projects, a good deal of research has been conducted at the *Musée de l'Homme* (Paris) to better describe the peopling of Central Asia. The main objective has been to characterize social and genetic differences, inferred through DNA testing, of populations living in Kyrgyzstan, Tajikistan, Uzbekistan and belonging to several ethnic groups like the Karakalpaks, Kazakhs, Kyrgyz, Tajiks, Turkmens, Uzbeks and Yagnobis.

We have shown (Martinez-Cruz et al, 2011) that these populations genetically cluster in two groups that closely follow the linguistic classification Indo-Iranian *versus* Turkic populations – with the exception of Turkmens, whose language is Turkic, but who cluster genetically with Indo-Iranian populations. This general distinction makes also sense in social and economic terms. On the one hand, the Turkic populations were traditionally herders while the Indo-Iranians were farmers, which

contributed to the emergence of a differentiation in some genes involved in food metabolism. For example, we observed some possible genetic differences between herders and farmers concerning the metabolism of milk (Heyer et al. 2011) and carbohydrates (Segurel et al. 2013), but not concerning the metabolism of proteins (Segurel, 2010). On the other hand social differences according to the group can also be observed. While both groups are patrilocal, Turkic populations are exogamous and inherit cultural traditions from the father, while Indo-Iranian populations are endogamous and both parents contribute to the transmission of culture and beliefs. These differences in social organization have influenced the genetic diversity (Chaix et al. 2004, Chaix et al. 2007, Segurel et al. 2008, Heyer et al. 2009) in a sex-specific manner: males from the Turkic populations show a low level of migration, while females conversely have a high level of inter-population migration.

During the genetic fieldwork we also probed the linguistic diversity of this region, as languages are widely regarded as a proxy for the cultural diversity among and between human populations (see Manni 2010 for a review). Actually, though human populations are overall genetically very similar, some heterogeneity can be found between geographically distant or ethnically different groups. It goes without saying that the noise and the randomness associated with a genetic sampling, limited in time and space, may reflect local phenomena that a more general classification of human groups would not show. Similarly, local linguistic differences may not fit well within general linguistic classifications that are often based on scholarly criteria culled from surveys, but not on general and replicable procedures. Therefore, in order to analyze the linguistic (cultural) diversity of the populations under study and to compare it to their genetic diversity, we recorded the language variants of the same individuals that agreed to donate their DNA. In this way and because we are working on dialectal variation, we did not have to rely on distant linguistic classifications that may not apply to the groups we approached (Baskakov 1966; Menges 1968; Andreev & Sunnik 1982; Johanson and Csató, 1998).

A genetic sampling of a given population is not immune from the confounding effect of recent migrations that make the inference about the ancestral genetic variability difficult. Bearing this limitation in mind, it may be misleading to compare a genetic classification of human populations with a linguistic phylogeny obtained without taking population contact into account. Historical linguistics,

largely based on the comparison of regular correspondences and on the comparative method—that is on the comparison of cognate words—focuses more on the language itself than on the speakers that kept it alive, speakers that are often non-native and bilingual. This is why we included loanwords in our linguistic analysis , i.e., because they are symptoms of population contact and admixture, i.e., the kind of phenomena that population genetics can address.

This article concerns the methodology we adopted to document local speech and to measure its differences in order to further compare it to the genetic diversity of the same populations. To do this we collected the realization of the concepts in the 200-word Swadesh list. First we identified the range of phonological variation of the two language groups under investigation (Turkic: Karakalpak, Kazakh, Kirghiz, Uzbek – Indo-Iranian: Tajik, Yagnobi), then we computed the aggregate linguistic differences among the different varieties using the Levenshtein (1966) algorithm (see also Heeringa 2004).

We should note finally, that our goal of understanding the human history of Central Asia will result in a description of its genetic variability, which will be summarized as a frequency matrix of given DNA motifs and consanguinity estimates. This makes the quantitative analysis of language variation, which results in matrices of linguistic distance, a most attractive corresponding goal, as it enables a statistical comparison with the genetics.

Table 1. Geographical distribution of the 88 respondents across 23 test sites. Latitude and longitude coordinates are expressed as decimal values

Language	Country	Region	Place	Latitude	Longitude	Questionnaires
Karakalpak	Uzbekistan	Karakalpakstan	Kokdarya	43,09	58,78	4
			Shege	43,77	59,02	4
Karakalpak			Halqabad	42,94	59,78	3
Kazakh			Raushan	43,04	58,84	4
		Bukhara	Gazli	40,08	63,56	7
Kyrgyz	Kyrgyzstan	Issyk Kul	Tamga	42,16	77,57	2
			Barskoom	42,17	77,64	
		Naryn	Kulanak	41,36	75,50	5
			Akmuz	41,25	76,00	3
Uzbek	Uzbekistan	Andizhan	Orday	40,77	72,31	4
			Soy Mahalla	40,77	72,31	3
		Bukhara	Zarmanak	39,73	64,27	3
			Novmetan	39,73	64,27	
		Karakalpakstan	Hitoy	43,04	58,84	3
	Tajikistan	Penjikent	Urtoqqishloq	39,49	67,54	4
Tajik	Uzbekistan	Bukhara	Zarmanak	39,73	64,27	4
			Novmetan	39,73	64,27	
		Fergana	Kaptarhona	40,25	71,87	5
			Rishtan	40,36	71,28	3
		Samarqand	Kamangaron	39,50	67,27	5
			Agalik	39,54	66,89	
	Tajikistan	Ayni	Shink	39,28	67,81	3
			Urmetan	39,44	68,26	4
		Penjikent	Nushor	39,11	70,86	3
		Tajikabad	Navdî	39,11	70,45	3
			Nimich	39,12	70,67	4
Yagnobi	Tajikistan	Dushanbe	Safedorak	38,57	68,78	3
			Dugova	Next to above	Next to above	2
						= 88

2. Methodology

2.1 Selection of linguistic test sites

Table 1 lists the locations where the linguistic data was collected. As the ultimate aim of the generalproject is to understand how cultural (linguistic) differences can influence human migration and geneflow, we deliberately selected test sites with an eye to the quite complex human and linguistic geography of this region of Central Asia. We have targeted populations living within the Indo-Iranian and Turkic-speaking zones but also at the borders between them (see Fig. 1). Further, when possible, we have documented linguistic varieties surrounded by a different language family. While several Indo-Iranian Tajik speaking groups live in the officially Turkic-speaking Uzbekistan, the opposite

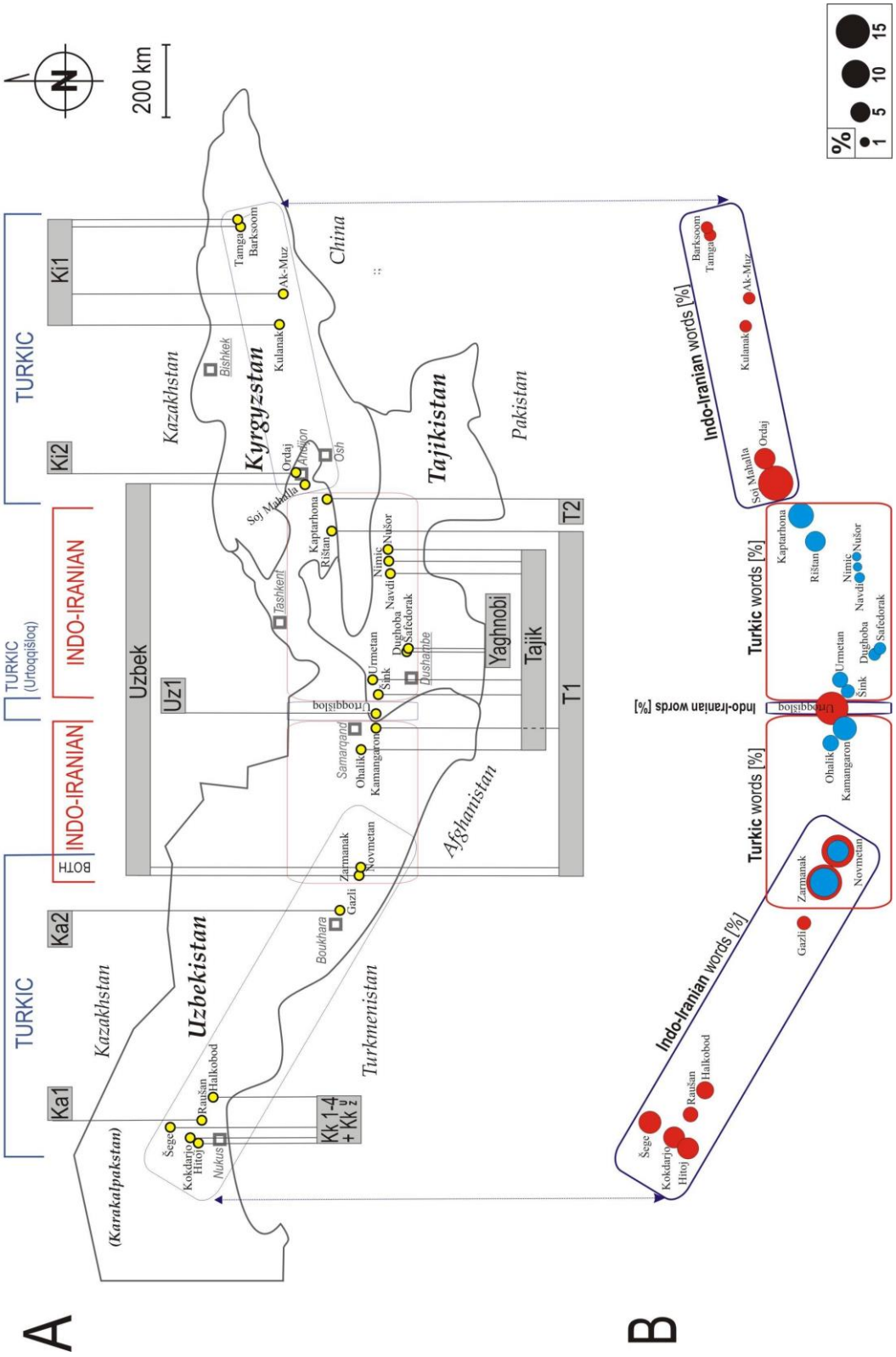

Figure 1. A. Geographical sketch of the region investigated. Test sites are reported in yellow. Major cities are reported as gray squares. Gray boxes with labels corresponding to the clusters found in the bootstrap tree of figure 2 are reported. Major linguistic classifications (Turkic, Indo-European) are reported at the top.
B. Test-sites in A have been plotted as circles whose surface is proportional to the percentage of loanwords from the other linguistic family appearing in the Swadesh list of 176 concepts (see scale on the right). Rounded-corner rectangles encompass the test sites belonging to a same linguistic affiliation and, within them, the loans from the other linguistic family are colored accordingly.

situation (Uzbek-speaking groups living in a Tajik-speaking area) is less common and only one village falls into this category (Urtoqqilsoq – See Fig. 1).

As far as the Karakalpak-speaking area is concerned, sampling sites were chosen according to the suggestions of a competent social anthropologist (Dr. Svetlana Jacquesson, personal communication).[2] More generally, suggestions about appropriate sampling sites came from local historians, social anthropologists, and local authorities. Our quest for "autochthonous" villages explains why the majority of the sites are away from large urban areas such as the cities of Tashkent (Uzbekistan) and Bishkek (Kyrgyzstan). An exception was made for the city of Bukhara, that has a special interest as it has been peopled since the Neolithic times.

For reasons related to the genetic ambitions of the larger project (explained in the introduction), the linguistic varieties under study correspond to populations representative of the two life-styles of Central Asia, farmers (i.e., the Tajiks) and pastoralists (the Kazakhs). While we wanted to investigate a larger region, linguistic varieties in the Tajik portion of the Pamir mountains have not been documented because of severe weather conditions during the fieldwork and because there was too little time to get used to the high altitude. There are no sampling sites in the southern region of Kyrgyzstan because local authorities did not allow investigation.

2.2 Linguistic inquiry

Local healthcare professionals selected the volunteers likely to be enrolled in the study after our arrival in the different villages, and our first contact with them was often the day of the DNA donation and, for some of them, the day of the linguistic interview. This is why we experimented with different approaches to identify the language variety spoken by the informants (Fig. 1). Before systematically adopting the extended Swadesh list of 200 words on which this paper is based, we experimented with using other word lists. The first one concerned "diagnostic" words chosen by F. Jacquesson (personal communication, see also Jacquesson 2002) meant to distinguish the speakers of three Turkic languages: Uzbek, Kazakh and Karakalpak. Because of the dialect *continuum* existing among Turkic

[2] See also Jacquesson Sv. 2002

languages of Central Asia[3] and the high degree of mutual intelligibility among the speakers of Uzbek, Kazakh and Karakalpak, we progressively adopted a longer list of diagnostic words according to Junusaliev (1966) and Menges (1968). While the results (not shown) were quite accurate and confirmed the linguistic literature, we felt that they were not portraying the dialect *continuum* appropriately. For this reason, we finally based our inquiry on the extended Swadesh word list of 200 items (Swadesh 1955, 1972).

2.2.1 Classification of spoken language varieties by using the extended Swadesh list.

Widely used in linguistic studies, the Swadesh list was developed for glottochronogy, i.e. dating linguistic events such as the splitting of the Germanic languages into North, East and West. It was designed to include the basic notions that we expect to find in the lexicons of all languages and to include, for comparative purposes, the words less likely to have been borrowed. Swadesh's approach was highly controversial, especially his notion of a "basic vocabulary". His idea was that essential words are more widely used and are more stable over time. For assaying linguistic relatedness, the words of the Swadesh list offer some advantages because speakers use them without hesitation, whilst more "marginal" words require longer reflection. Finally, the essential words are generally simple, not derived. Many articles have addressed the effectiveness and appropriateness of the Swadesh list and we redirect interested readers to reviews (Kessler 2001; McMahon & McMahon 2005; Holman et al. 2008).

While the standard Swadesh list seems a convenient base to collect phonological and lexical data and to compute linguistic distances between the speakers, we had to adapt it to this specific linguistic context by excluding some words (see Supplementary Materials, Tab. S1). To minimize polysemy, a better list had to be developed and used instead, but our fundamental choice to use the Swadesh list (because of its wide use in linguistics) was modified only slightly. By using the Swadesh

[3] This is particularly true for the Kyrgyz, the Kazakhh and the Karakalpak, and partly for the Uzbek of Karakalpakstan. The official Uzbek, which has been influenced by Iranian languages, is slightly different, although its Turkic basis still shimmers through, further attesting to the phonetic continuum among the Turkic languages.

list of concepts we intend to facilitate the comparison with linguistic data collected in the same format in other populations.

2.3 Informants, protocol and linguistic database

The Swadesh 200-word list was submitted to 88 people during fieldwork in Uzbekistan, Kyrgyzstan and Tajikistan from 2003 to 2007 (Tab. 1). More than 17,000 items have been digitally recorded and manually transcribed (Supplementary Materials, Tab. S2). All informers were approached by the same linguist (PM), usually in rural health centres. For genetic-testing purposes, they generally were male adults aged of at least 40 years. If the social background of the informants is uneven (mainly manual workers in rural areas and low middle-class or middle-class in urban areas), they all went to school during the times of the Union of Soviet Socialist Republics (USSR) and could understand Russian well.

In practice, the informants were asked to orally translate into their mother tongue the Swadesh words that were asked in Russian.[4] The conditions of the linguistic interview (microphone on the table, empty room, medical environment, interview conducted in Russian by a linguist coming from Paris, specific request to speak clearly, words pronounced out of context) could not have been more formal, meaning that recorded variants are quite far from natural language conditions and inter-individual variability could not be captured fully. But, unfortunately, no alternative setup was possible.

As we were unable to identify a priori *the* speaker best representing the variety spoken in a given site, we adopted a sociolinguistic approach and collected about four questionnaires per village (though their number varied according to the size of the villages and was sometimes conditioned by

[4] While informants were asked to speak their everyday language, some made visible efforts to "speak well", reproducing the official language. We stressed that we were interested in recording their traditional language and history, which was often a convincing argument in the context of the nationalistic policies pursued by many Republics after the collapse of the Soviet Union and their independence. We note that the Uzbek spoken in Uzbekistan is much more homogenous than the Tajik of the Tajik-speaking minorities living in Uzbekistan The speakers who agreed to "play the game" and were proud to use their everyday language are rare; they are notably the informants coded Ka-Gazli 5, Ka-Gazli 7, T-Kamangaron 2, T-Kaptarhona 5, T-Navdî 1, T-Nimich 3, T-Nimich 4, T-Nushôr 3, T-Rishtan 3, T-Urmetan 1, T-Urmetan 3. It is significant that these are essentially the Tajiks of Uzbekistan, less influenced by the Tajik norm. For example, they are perfectly aware that they employ the syllabic consonants [Ç] instead of groups [V+C].

the lack of Russian speaking volunteers -- see table 1). The number of questionnaires concerning Yagnobi speakers is lower than the average because of difficult working conditions during the inquiry.

When a respondent did not understand a concept (usually one used less in everyday speech), a drawing corresponding to the concept was sometimes shown. When the informant provided two different responses, we have generally chosen the first realization because the second one was usually an attempt to speak closer to the norm. We note that words were asked in the order established by Swadesh, except for the first 21 (abstract words in a grammatical context) that were asked at the end (Supplementary Materials, Tab. S1). The whole linguistic interview lasted for about half an hour for each informant.

To be sure that no misunderstandings arose during the interview and that recorded words indeed corresponded to the concepts of the Swadesh list, all realizations have been verified according to several references (Andreev and Sunik 1982, Balci B. et al. 2001, Baskakov 1966, Junusaliev 1966, Kerimova 1959, Moukhtor et al. 2003, Rastogueva 1963, 1964). Later, we transcribed the realizations in IPA (International Phonetic Alphabet) but we did not try to reconstitute the phonology. Each transcription has been compared to the corresponding recording several times. The phonetic transcriptions were subsequently translated into the X-Sampa codification for computational processing (Wells 1997, for more details please see the internet address http://coral.lili.uni-bielefeld.de/LangDoc/EGA/Formats/Sampa/sampa.html). It must be pointed out that we are dealing with the phonetics of words pronounced in isolation. For example, the devoicing of the final consonants in all the languages of the region studied does not occur before a consonant in the following word or in a suffix.

2.4 Computational analysis

We estimated the similarity of varieties using a pronunciation distance metric, in other words, ignoring syntax and morphology. Since lexical differences also result in pronunciation differences, these are incorporated into the method. It is occasionally objected that behavioral tests are needed to determine how closely related language varieties are, e.g., tests of (mutual) intelligibility, and the objection is not without merit. But behavioral tests, e.g., of intelligibility are not only expensive to conduct, but also

reflect linguistic similarity only partially, as language attitudes and experience likewise play a role. Finally, as Gooskens et al. (2008) show, pronunciation similarity together with lexical overlap (shared cognates) predicts intelligibility quite well (explaining 81% of variance).

We therefore compared the phonetic transcriptions of the pronunciations using the Levenshtein algorithm, also known as Edit Distance (Levenshtein 1966). When calculating edit distance between a pair of words in two different varieties, the algorithm seeks for the minimal set of operations that can be used to transform one realization into another. The operations can be insertions, deletions, substitutions or swaps and each is associated with a cost (Tab. 2).

Table 2 Levenshtein distance example concerning a pairwise distance computation between two realizations for the word 'round' in two languages. Tajik horizontal (Agalik) and Uzbek vertical (Zarmanak). The algorithm begins with all the cells in the matrix empty except for the zero in the upper left hand corner. Each cell is then filled in with the minimum of three possible values: (i) the value in the cell to the left plus one (corresponding to an insertion); (ii) the value of the cell above plus one (corresponding to a deletion); or (iii) the value of the cell diagonally above and to the left plus one if the row and column indices differ, or plus zero if they are the same. The cell (l,l) involved a null change with respect to the cell (j,0). The value in the lower right-hand corner (4) is then the minimal edit-distance between the two strings, the least number of edit operations required to transform *ajlana* to *lUnda*. See Tab. 1 for more geographical details.

		l	**U**	**n**	**d**	**a**	
	0	1	2	3	4	5	*null*
a	1	1	2	3	4	4	*insertion*
j	2	2	2	3	4	5	*insertion*
l	3	2	3	3	4	5	*null*
a	4	3	3	4	4	4	*substitution*
n	5	4	4	3	4	5	*deletion*
a	6	5	5	4	4	4	*null*
						4	

Although we have experimented with elaborate cost schemes, we have generally found simple schemes to function effectively when the purpose is to characterize the overall similarity among varieties (Heeringa 2004: page 186). Therefore, a standard cost scheme in which all operations cost a single unit (1.0) is adopted, here, in Levenshtein distance computation. Heeringa (2004) presents the application of Levenshtein distance in great detail. We ensure, roughly, that only vowels substitute for

vowels, and consonants for consonants, and the distance scores are normalized according to word-length (see Heeringa et al. 2006 for details). The guarantee is only rough because we do allow vowels to substitute for sonorous consonants (/r, l, n, m/ etc.) and approximants such as /w, j/ may substitute for vowels but also for consonants.

As in previous dialectometric research conducted by the Groningen group, the software package L04, developed by Peter Kleiweg (http://www.let.rug.nl/kleiweg/L04) and GabMap (www.gabmap.nl/) (Nerbonne et al. 2011), are adopted for analysis. These packages contain several methods to analyze phonological and lexical data statistically, building on the Levenshtein measure for string data, but including routines to analyze numerical data such as frequencies formant values, and others to analyze categorical data (such as lexical choices). The focus is on analyzing string data such as IPA transcriptions.[5]

For the purpose of testing effectiveness in distinguishing different varieties, L04 is used to calculate a distance between each pair of words, and then an aggregate distance score for each pair of sites (the mean of word distances). We collected the aggregate site distances into a site × site matrix, which was further analyzed using multivariate analyses and hierarchical clustering.

2.5 Matrix Generation

Given our use of Levenshtein distance as a measure of pronunciation difference, it is natural to continue using distance-based methods as opposed to character-based methods to understand how the linguistic varieties relate to one another. Kassian (2015) confirms the general wisdom of preferring distance-based analyses. Site × site matrices of mean edit-distances were generated for the entire dataset and for the two main language groups in it, namely Turkic (Karakalpak, Kazakh, Kyrgyz and Uzbek) and Indo-Iranian (Tajik and Yagnobi). Though the Yagnobi language is classified in a different subgroup (Eastern Iranian) than Tajik (Western Iranian), we processed them together.

[5] Naturally there are alternatives available, notably the AJSP work and Jäger's (2013) work, both cited in the introduction, but also Mattis List's LingPy programs (List 2014). In fact we have developed more sensitive measures as well (Wieling, Margarethe and Nerbonne 2012). Given our focus on varietal level comparison, we feel that more sensitive measures are unlikely to contribute much.

We also processed the dataset per wordlist, meaning that linguistic distances between all the pairs of speakers have been computed according to the shorter (100 words) and to the longer Swadesh list (200 words). There is a discussion in the literature as to whether the 100-word or 200-word list is better for the purpose of assaying linguistic relatedness, and we wished to know the degree to which the relatedness would overlap depending on which of the two sets was used. Because the 100-wd set is largely a subset of the 200-wd set, the two are not at all statistically independent, so we will not attempt to interpret the significance of the (very high) correlation we obtain.

2.6 Relations among varieties

We investigate the structure of the site × site matrix of linguistic distances using (bootstrap) clustering (Nerbonne et al. 2008) on the one hand and multi-dimensional scaling (MDS) on the other (Nerbonne, Heeringa and Kleiweg 1999). We do not wish to assume that the varieties are tree structured, i.e., the result of purely vertical inheritance with occasional splits. The high level of contact, systematic migration and potential for population admixture we find in Central Asia suggests that we should expect to find horizontal transfer as well. We therefore prefer techniques such as MDS, at least initially, to phylogenetic inference, which does assume a tree structure. We hasten to add that we have no reason to doubt the clear separation of the Turkic from the Indo-Iranian varieties. But, as we note below (4.4.4), the frequency of lexical borrowing confirms our suspicion that horizontal transfer was also an important factor determining the current relations among the varieties studied. This will be reflected in MDS plots but not in phylogenetic trees.

The linguistic distance matrices (between all pairs of speakers and between all pairs of speakers within a language group) have therefore been analyzed and visualized as a consensus bootstrap tree (Fig. 2) and as classical multidimensional-scaling plots where the squared error is minimized (Fig. 3). The bootstrap tree guards against too strong an assumption of tree-like structure by resampling original data (with replacement) a hundred times obtaining 100 randomly resampled new datasets containing the same number of items (words) as the original (though with some items appearing repeatedly and some not appearing at all, due to the randomness of the resampling procedure). More details about the procedure can be found in Nerbonne et al. (2008). The length of a

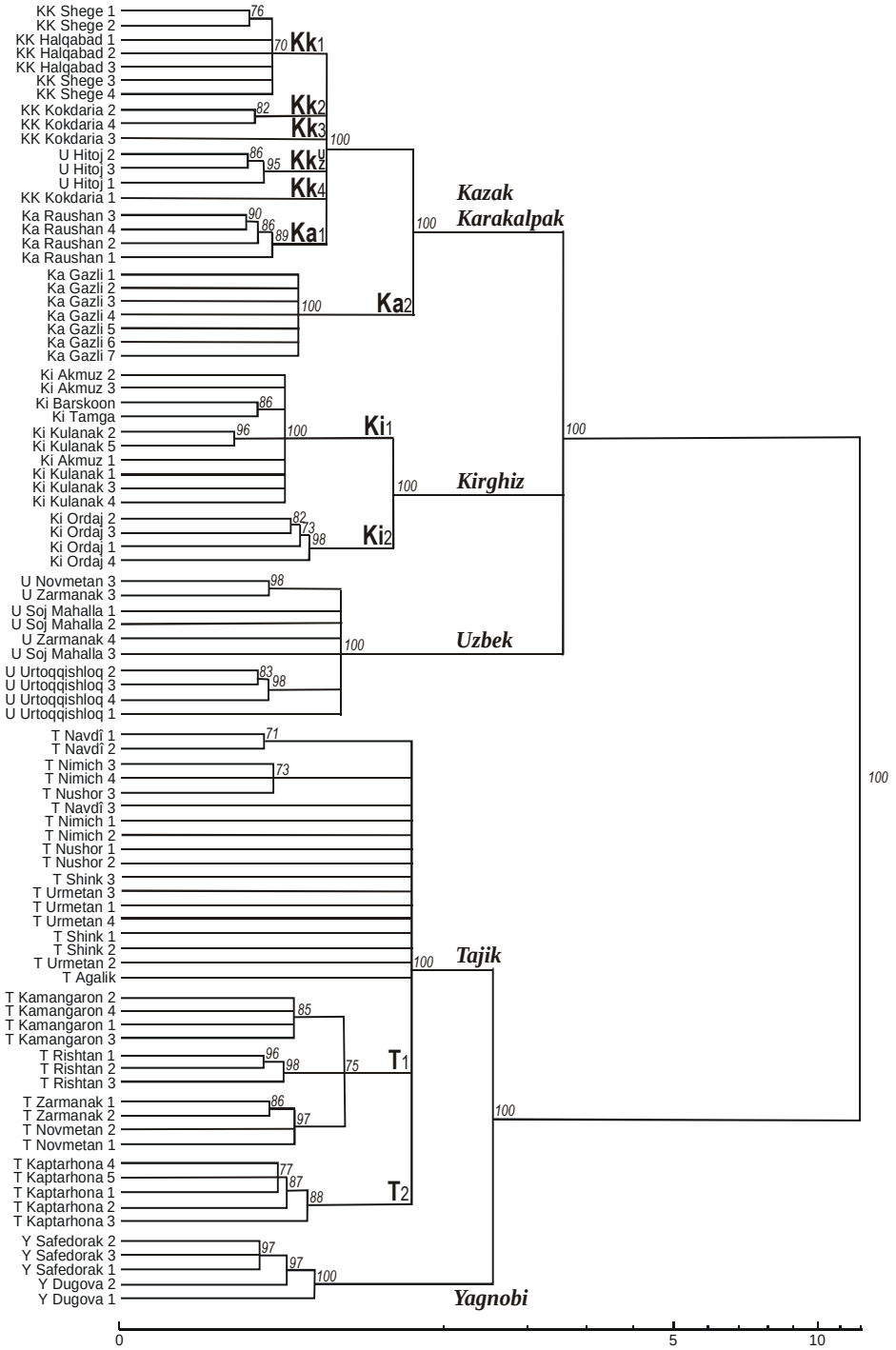

Figure 2. Bootstrap consensus tree accounting for the linguistic similarities and differences between 88 informants interviewed in 23 test sites according to Levenshtein linguistic distances. The scores at each node of the consensus tree correspond to the number of times each bifurcation is observed in the 100 trees obtained from the 100 matrices corresponding to 100 datasets re-sampled from the original dataset with the bootstrap method. Nodes not supported by at least 70% of the bootstrap re-sampled 100 datasets have been collapsed, thus giving sometimes rise to a "comb" geometry meaning that no robust hierarchal clustering can be assessed at the corresponding level of the tree. This cut-off value is arbitrary, though quite standard in similar analyses. Major linguistic bifurcations are very stable (bootstrap score = 100). Further clusters are labeled by an alphanumeric code (ex. Kk1, Kk2, Kk3, etc.) and also reported in figure 1 for visual ease in geographical comparisons. Distance matrices concern aggregated Levenshtein distances accounting for pair-wise comparisons of the realizations of 176 words that are included in the 200 list of Swadesh concepts (see Supplementary Materials table S1 for details about the wordlist).

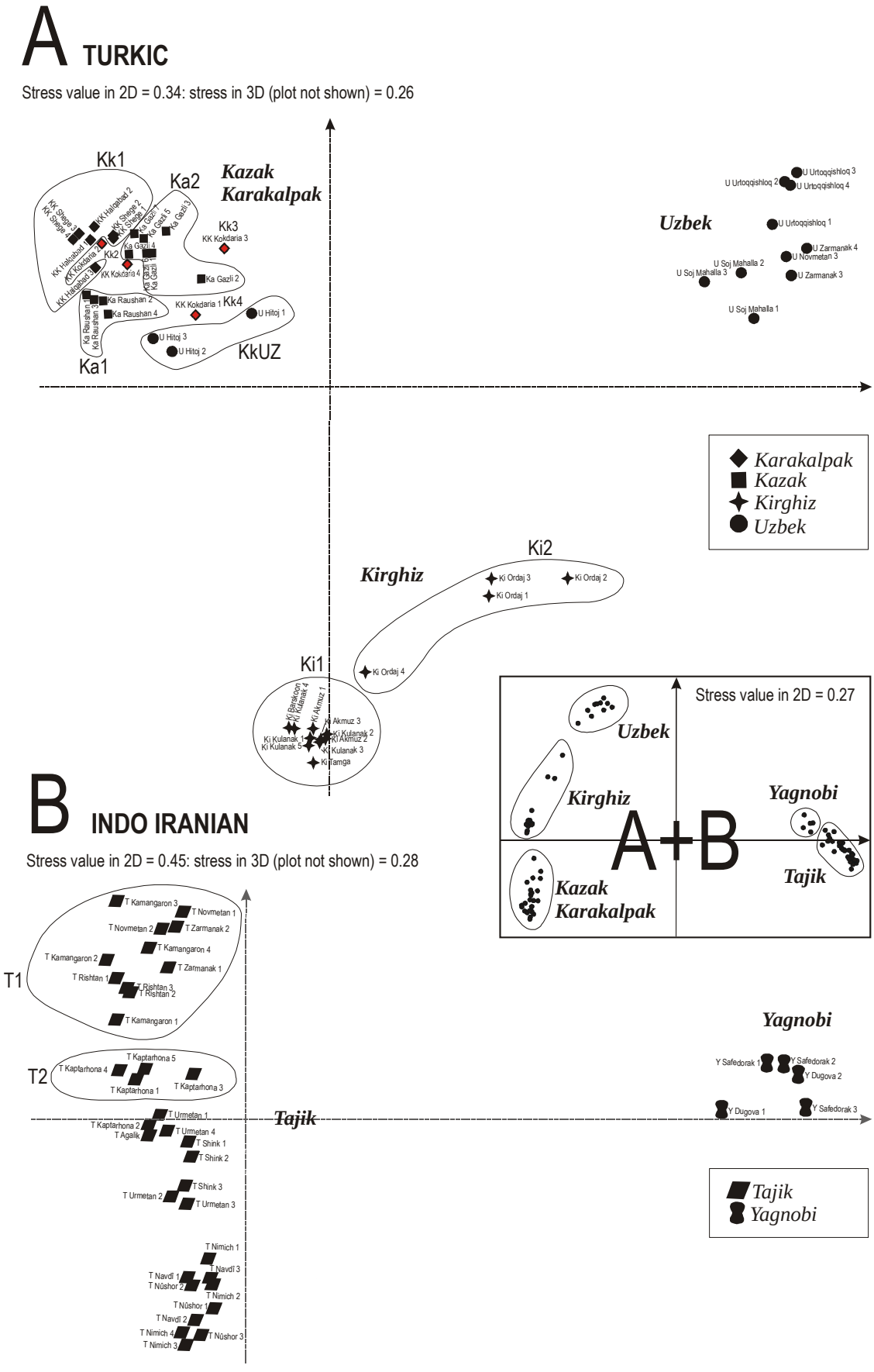

Figure 3. Two-dimensional multidimensional scaling (MDS) plots accounting for the Levenshtein linguistic distances between the 88 informants interviewed in 23 test sites. Informants are displayed by linguistic family

and altogether (A. Turkic; B. Indo-European, A+B altogether). Symbols are provided when necessary to distinguish single languages. Red-filled symbols correspond to samples that stand out in a three-dimensional representation (not shown). This is the case of the village of Kokdaria in A. and of the village of Kaptarhona in B. All plots provide a representation of variability that is somewhat complementary to the tree displayed in figure 2, though they are not based on re-sampled matrices. For cross-comparison ease, some groups of points are encompassed by a circle that corresponds to the clusters (ex. Kk1, Ka1, Kk2, etc.) appearing in figure 2 and also in figure 1. Stress values, corresponding to the deformation of each projection, are reported for each plot both for two-dimensional analysis (shown) and for three-dimensional one (not shown).

branch reflects the cophenetic distance from an internal node to the daughter nodes, which may be leaves (sites). The robustness of the clustering (see scores at each node of the tree in Fig. 2) is proportional to the number of times a cluster appears in the different 100 trees. In Figure 2 we set a cutoff value of 70%, meaning that all nodes supported by fewer than 70 of 100 iterations were collapsed. 70% is an arbitrary threshold commonly accepted as a reasonable compromise. The many non-binary branches in Fig. 2 (see the Tadjik leaves as well the second Kazakh node) reflect groups whether further tree-like structure could not be reliably ascertained. While we used three different clustering algorithms, with results that are largely comparable, the clustering method we present was produced by Ward's method (Fig. 2). Ward's method is one of the four techniques that have been found to recognize hierarchically organized groups in dialects well (Prokić and Nerbonne 2008).

The major clusters in the bootstrap tree (Fig. 2) have been labeled and some labels are reported in figures 1 (part A) and 3 (part A and B) for the readers' visual ease.

2.7 Loan word detection

To determine the number of loans, we followed a three-step procedure. First, we filtered out all the speaker pairs from the same language group (i.e., both from the Indo-Iranian group or both from the Turkic group) because we focus on the borrowings that have occurred from one language family to the other. Second, for each pair of speakers from different groups, we used the Levenshtein algorithm to compare the transcriptions of each word probed. Our leading hypothesis was that near-identical pronunciations of the same word in different language families would indicate that the word had been borrowed from one language family into the other. The third step consisted in evaluating this hypothesis against PM's expert judgment as to which words were borrowings. For this we used a technique from information retrieval, the 11-pt interpolated average precision curve (Manning,

Raghavan and Schütze 2008:145-148), which compares the Levenshtein scores to PM's expert classification of words into borrowed and un-borrowed (Supplementary Materials, Tab S2) We elaborate on this below.

For each concept in the Swadesh list, and for each pair of sites, we obtained a pronunciation distance—the edit distance between the pronunciations realized at the one site from the pronunciations realized at the other. We use these single-word distances to detect likely loan words, under the leading hypothesis that words from unrelated language families that are very similar probably are loan words. We quantified the success of recognizing loan words using precision and recall (Manning et al. 2008). Recall is the fraction of genuine loanwords that is correctly recognized, i.e., the percentage of realization-pairs expertly classified as loanwords which are also automatically recognized as such (i.e. by having a low score for edit-distance). Precision, the fraction of true positives, is the percentage of the pairs identified as loanwords on the basis of low edit-distance scores, which were also expertly classified by PM as loanwords. Note that there is an obvious trade-off between precision and recall: the lower we set the edit-distance threshold, the better our precision gets, while recall, however, drops. For this reason, we prefer to examine a curve, and one conventional presentation graphs the average precision at eleven different recall levels, namely 0%, 10%, etc. through 100%. Fig. 4 presents our detection of loan words as the curve showing precision at these eleven different levels of recall. It shows that precision is nearly perfect at low edit distances, while rec all is still 50%. See Fig. 4.

Finally, we also investigated whether the words related by loan as a set differ from other words (whether the mean realization differences differ significantly), and we tested whether the distribution of edit distances might be better understood as a mix of two distributions, using the EM algorithm (Ju 2002), implemented in the 'mixdist'-package in R (http://www.r-project.org/). This routine tries to analyze an input distribution as the sum of two Gaussians. The results may be examined in Van der Ark et al. (2007). This confirmed the cutoff point suggested by the precision-recall analysis.

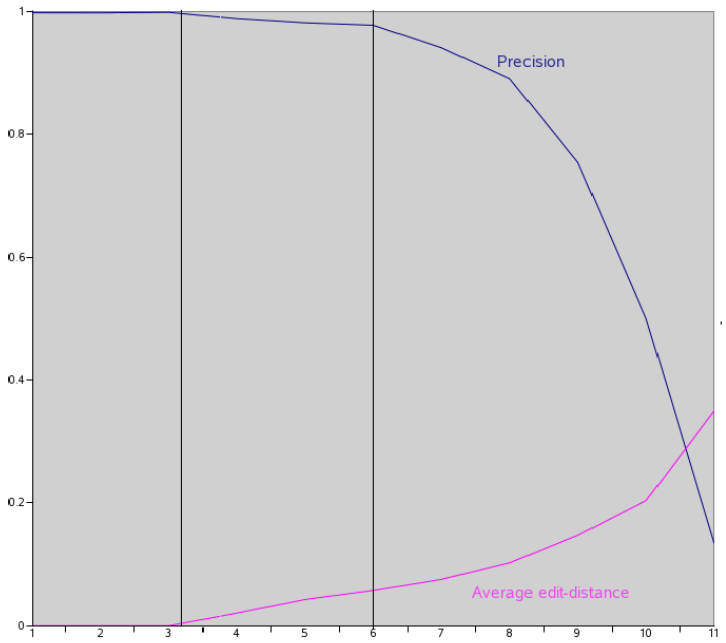

Figure 4. The precision (or accuracy) of loan word detection as a function of the recall (the fraction of loanwords detected). Recall increases as the Levenshtein distance threshold drops. See text for further explanation.

Before leaving this section we would like to note that the reliable detection of loan words is a further task that might be assigned to edit distance approaches to dialectology and diachronic linguistics. We are certain that the fairly rough approach taken here can be improved, for example using more sensitive measures and perhaps also by exploring the sensitivity of borrowed words to the structural disparities between their source and target languages.

3. Results

3.1 General sketch of phonetical variability

It would be vain to try to establish, on the basis of a Swadesh list of 176 terms (24 words were excluded – see Supplementary Materials, Tab S1), the regular connections between languages of the same family. In this section, our purpose is to highlight the phonetical similarity and differences of different Central Asian varieties to suggest that their diversity falls in a range of diversity comparable to the European dialects we have studied so far (Gooskens and Heeringa 2004; Nerbonne and Siedle 2005; Wieling et al. 2007; Prokić et al. 2009; Wieling et al. , 2013; Šimičić et al. 2013; Montemagni et

al. 2013). As a consequence the computational methods we used to measure the linguistic diversity, originally designed to analyze dialect diversity in Europe, can be regarded as appropriate tools for the task at hand. The linguistic variation we find is within the bounds we find in dialectological studies, where the tools have been found to validly detect the relations among varieties (Heeringa et al. 2002, 2006). See, too, next section.

3.1.1 Turkic languages

There are recurrent phenomena in the Turkic languages (Karakalpak, Kazakh, Kirghiz, Uzbek) of this region of Central Asia: (1) the devoicing of the final consonants (words pronounced singly) (ex: *muz̥ / mʏz̥* 'ice'); (2) the frequent consonant palatalization before [e/ɪ] (ex: *bʲes / bʲɪs* 'five', *kʲeŋ* 'wide'); (3) in Kazakh and Karakalpak, the labialization of plosives before [ɵ] and [u], and the epenthesis of [w] in word-initial position (ex *tʷɵrt* 'four', *kʷɵz* 'eye'; *urman* > *wʊrman* 'forest'); (4) in Kazakh and Uzbek, the deletion of interconsonantal [i], with the subsequent assibilation of the next consonant (ex: *qɪsqa* 'short' > *qʂqa*); (5) in Uzbek, the frequent deletion of [u] (ex: *tuxum* 'egg' > *txum*); (6) in Kazakh and Uzbek, the velarization of final [l] (ex: Kaz. *qɔɬ* 'hand'; Uzb. *kwʊɬ* 'lake'); (7) in Kazakh, the tendency to weakening of initial [h] (ex: *hajal* > *ajal* 'woman'); (8) in the Kazakh variety spoken in the village of Shege (Fig. 1) and the Karakalpak variety spoken in the village of Halqabad (see Fig. 1), the lenition (voicing) of initial [t] (ex: *tuman* 'frog' > *duman* ; *tuz* 'salt' > *duz̥*). Some more phonetic tendencies in the Turkic languages are listed in table 3.

Table 3 Sketch of phonetic tendencies in the Turkic languages.

Kyrgyz	Kazakh	Karakalpak	Uzbek	Commentaries
a	a		ɔ	Vowel correspondences.
y	ʏ		ʏ/u	
i	ʏ		e/ɪ	
#je	#ji		#i	
ɔː	aw		aɣ/ɔɣ	Deletion of intervocalic consonant and vowel lengthening in Kyrgyz.
te	tʲe			Consonant palatalization before [e].
#p	#b		#b[6]	Consonant correspondences in initial position. Tendency to sonorization in Karakalpak. Palatalization of dentals before [e].
#m	#m		#b	
#te	#dʲe	#tʲe	#dʲe	
#t	#t / #d	#d	#t	
#ʃ	#ʃ	#s	#ʃ	Assibilation of initial [ʃ] in Karakalpak.
t͡ʃ#	ʃ#		t͡ʃ#	Lenition of final [t͡ʃ] in Kazakh et Karakalpak.
ʃ#	s#		ʃ#	Assibilation of final [ʃ] in Kazakh et Karakalpak.
#d͡ʒ	#ʒ		#j	Lenition of [d͡ʒ] in Kazakh, Karakalpak et Uzbek.
#k	#g		#k/t͡ʃ	Examples of correspondences for velars.
#g	#k		#k	

3.1.2 Iranian languages

In recorded Iranian varieties the changes are mostly lexical. Nevertheless, we observe phenomena similar to those occurring in Turkic languages, that probably shows an areal influence like the palatalization before [e] (ex: Tjk *sʲetɐ* 'three') or the deletion of intervocalic [u] with the subsequent consonant syllabization (ex: Tjk *tuxum* 'egg' > *txum / tuxm̩ / txm̩ / tx̩m* ; *dum* 'tail' > *dm̩*). For Yagnobi varieties, we do not have enough informers to establish fine comparisons.

3.2 Measures of the linguistic variability

3.2.1 Matrix consistency

Distance matrices were generated for the entire data collection and, separately, by language family (Turkic vs. Indo-Iranian sites). Consistency measures (Cronbach's α) confirm the strong signal in the data that does not diminish significantly when the two families of languages are merged in the analysis (Tab. 4). We also computed distances by word list, that is according to the shorter 100-word Swadesh

[6] Uzbeks of Soy Mahalla: #p.

list or according to the longer 200-words Swadesh list, and in this case matrix consistency is also very high (Tab. 4).

Table 4. Correlation coefficients for multidimensional scaling (MDS) analysis in three dimensions. The relations among the Indo-Iranian group are a bit more difficult to approximate in three dimensions, probably because Tajik and Yagnobi show more complex pronunciation differences .

Word Subset	*All Respondents*	*Turkic*	*Indo-Iranian*
All Words	0.984	0.981	0.967
Swadesh-200	0.982	0.981	0.967
Swadesh-100	0.987	0.981	0.968

3.2.2 Representation of variability, the bootstrap clustering

The consensus bootstrap clustering of figure 2, displaying all the 88 informants we approached during the fieldwork, shows a major split between the Turkic languages in the top half of the diagram and the Indo-Iranian languages in the bottom half. The clustering provided by this consensus bootstrap tree is quite robust because all major nodes are supported by a score of at least 70% (100% for major clusters).

Concerning the Turkic group, there are three major clusters (groups underlined correspond to clusters in Fig. 2): Kazakh/Karakalpak, Kyrgyz and Uzbek. The Kazakh/Karakalpak cluster is made of two sub-clusters. On the one hand we have several Karakalpak varieties (Kk1 mainly corresponding to the village of Shege; Kk2-Kk3-Kk4 corresponding to Kokdaria), the Kazakh variety spoken in the village of Raushan (Ka1) and the variety spoken in Hitoj corresponding to a population identifying itself as Uzbek though we classify it as Karakalpack (Ka^{UZ}). On the other hand there is the cluster formed by the Kazakh speakers of Gazli (Ka2). We note that the two Kazakh varieties in our dataset (Raushan and Gazli) do not form a single cluster by themselves, Raushan being closer to Karakalpak than Gazli is. The Kyrgyz cluster is divided into two subclusters, one corresponding to all the varieties spoken in Kirghizstan (Ki1) and a second consisting of the four speakers of Orday (Ki2), a village that is now part of Uzbekistan. If the Uzbek group is quite homogeneous (villages of Novmetan, Zarmanak and Soj Mahalla), we note that three speakers of Urtoqqishloq are grouped in a subcluster (Uz1), which makes sense given the isolated position of this village in the linguistic landscape of the region (Fig. 1, part A).

The Indo-Iranian cluster is split into two groups that correspond to the two languages it includes: Tajik and Yagnobi. While the Yagnobi cluster shows some differences between the two villages of Safedorak and Dugova, which are geographically very close to each other, the Tajik cluster is more complex. Even if two speakers from Navdi (Navdi1 and Navdi2) and three speakers form Nimich and Nushor (Nimich3, Nimich4, Nushor3) are grouped together, we still note that other speakers from these villages belong to independent branches, as do the speakers from the villages of Agalik, Shink and Urmetan. A closer look at figure 1 lets us recall that these villages are in Tajikistan (labeled as 'Tajik' in Fig. 1, part A), apart from the village of Agalik, which is in Uzbekistan, although it is not far from the border. Their belonging to single branches indicates considerable inter-individual variation that is unexpected seeing that Navdi, Nimic and Nushor are very close to each other. In addition, there are two subclusters within the Tajik cluster, one corresponding to the village of Kaptarhona (T2) and another to the villages of Kamangaron: Rishtan and Zarmanak/Novmetan (T1). These five Tajik-speaking villages (T1 and T2) are located outside Tajikistan (in Uzbekistan), which probably explains their clustering. We note that Zarmanak and Novmetan are very close and host a bilingual community, though speakers use only one language (Tajik or Uzbek) at home; this explains why speakers from Zarmanak and Novmetan appear in different clusters.

3.2.2 Representation of variability, The Multidimensional Scaling (MDS) clustering

The MDS analysis (Fig. 3) is complementary to the hierarchical analysis of the boostrap tree (Fig. 2) and visually shows the extent of the linguistic differences we measured. The plot that concerns both language families (Fig. 3 'A+B') shows that the variability within the Turkic languages is much higher than the one within the Indo-Iranian family, since Kazah/Karakalpak, Kighiz and Uzbek varieties span a bigger surface of the plot than the one occupied by Tajik and Yagnobi speakers, who are much closer to each other. In general, we note that each language—with the exception of Kazakh and Karakalpak that form a single swarm of points—corresponds to a non-overlapping cluster, confirming the major groups of the bootstrap tree (Fig. 2). The large linguistic distances between the speakers of the two language groups, Turkic and Indo-Iranian, distort the representation and muddy the topology of the points corresponding to the same languages, or the same group of languages. This is why we have computed separate plots for Turkic (Fig. 3, part A) and Indo-Iranian speakers (Fig. 3, part B).

Concerning the Turkic group (Fig. 3, part A), the Uzbek and Kyrgyz and Kazakh/Karakalpak speakers are nicely separated in three non-overlapping swarms of points. In more detail, the three Uzbek speakers of Urtoqqisloq, Tajikstan (they correspond to the Uz1 cluster of the bootstrap tree of Fig. 2) are next to each other and slightly farther from the other Uzbek speakers. The Kyrgyz of Ordaj (cluster Ki2) are quite distinct from the other Kyrgyz living in Kyrgyzstan (besides the speaker Ordaj 4). We note that the topology of the Kazakh/Karakalpak speakers is more complex and does not correspond well to the classification of the bootstrap tree as the clusters Kk1, Ka1, Ka2 and Kk^{UZ} are not distinct from each other in the MDS plot. This phenomenon is also related to the distortion of the two-dimensional presentation of the plot, and in fact a closer look at the third dimension (not shown) provides evidence for the separate position of the Ka2 cluster corresponding to the village of Gazli and for the considerable linguistic heterogeneity within the village of Kokdaria (red diamonds).

The plot involving Indo-Iranian (Fig. 3, part B) provides clear evidence of the difference between speakers of Tajik and of Yagnobi. The latter language is nowadays endangered and only spoken by a small community. Though inter-individual diversity (idiolects) is decreasing (this is what was observed during the fieldwork), because speakers are in the process of being integrated into the Tajik group with a loss of linguistic diversity, the linguistic differences based on the Swadesh list still appear to be substantial. As far as Tajik speakers are concerned, the two sub-clusters T1 and T2 highlighted in the tree re-appear here clearly, though not as distinctly as the bootstrap tree would suggest. This is probably related to a lack of accuracy in the two-dimensional representation that is linked to a stress value quite high (0.45), but not far from those predicted according to the tables of Sturrock and Rocha (2000) with 88 objects (stress = 0.39). A closer look at the third dimension shows a clear separation between the five speakers from the village of Kaptarhona and all the others. The Tajik speakers of the villages of Nimich, Navdi and Nushor are linguistically similar, as expected given their geographical vicinity, while those from Agalik, Shink, and Urmetan are more varied. Actually, the tree suggests considerable individual variation within all the six villages of Nimich/Navdi/Nushor and Agalik/Shink/Urmetan, which is reflected in the relatively larger cophenetic distances (branch lengths) in the dendrogram. The MDS plot in figure 2, part B, also represents this group as fairly diverse (see lower left-hand quadrant of the plot). The reason is related

to the bootstrap procedure that seeks for a consensus tree over resampled datasets, whereas the MDS plot concerns the full dataset without any resampling. Otherwise, both methods (boostrap and MDS) point to the significant linguistic heterogeneity among the speakers of the villages of Agalik, Shink and Urmetan.

3.2.3 Comparison of results using100-wd vs.200-wdSwadesh list.

To gauge the possible impact on the results of the two different Swadesh lists, we computed an overall Levenshtein distance matrix based on the shorter list (100 words) and another on the full list (200 words). The two matrices are almost identical (Mantel test correlation: 0.997***) and the comparison of the topology of samples in MDS plots (both highly correlated to original distance matrices r = 0.90 (Indo-Iranian) and r=0.94 (Turkic) plots not shown) is almost identical. We just noted an increased distance of the Yagnobi from the Tajik speakers in the reduction based on the 100-wd. sample. This result is in agreement with the claims that the shorter Swadesh list is more conservative and therefore more likely to reflect older linguistic relations, but it should be clear that the differences are minimal. Because the concepts it contains are used very frequently, they are therefore less likely to be borrowed (Kessler 2001, McMahon and McMahon 2005).

3.3 Loan word detection

In the Precision/Recall analysis (P/R), about a half of the full set of pronunciation-pairs are considered (about 250,000 pairs involving an Indo-Iranian speaker and a Turkic speakers). We reach 100% recall at after roughly 37.000-thousand records, at which point the last pair which had been classified as belonging to the same cognate group (and therefore as a loan word in one of the languages) is found. Therefore the precision vs. recall (P/R) curves of figure 4 are based on about 15% of the records. Fig. 4 shows that, initially and up to the thirtieth percentile, precision is almost perfect, meaning that all the words up to this point correspond to those manually classified as cognate. Since the pair of realizations is found in two different language families, the pair consists of a loanword and its cognate "source" in another language family. We realize that we are using the term "cognate" loosely here to include borrowing; in this sense English 'beef' and French *bœuf* are cognate, and, indeed they arise from the

same source, the English word arising via borrowing from French. It makes sense that word pairs with very low edit distances would be borrowings, and the P/R analysis starts by considering the lowest edit-distances (zero-linguistic distance = identical pronunciation).

The analysis shows that recall edit-distances are close to zero up to the thirtieth percentile. After the fiftieth percentile the precision-score starts dropping more dramatically, which happens at an average Levenshtein-distance of about 0.06. Based on this we can infer the score corresponding to edit-distances low enough to detect a loan reliably. Two thresholds were chosen. The first one considered is the 0.06 normalized edit-distance, based on a precision score of 0.977 at the fiftieth percentile, that is, just before the precision drop. The second threshold is the 0.02 score, at the thirtieth percentile, up to which precision scores are almost perfect. For this second threshold it can safely be stated that all the pairs that are below it can be considered loans. Once the thresholds are defined, no more runs of the P/R-analysis are necessary. We add that we also tried applying the P/R-analysis within the same language groups (Turkic or Indo-Iranian) but, within each group, the degree of cognacy is too high to identify possible loans that are hardly distinguishable from cognates pairs. We thus cannot detect what linguists call "intimate borrowing" (Jeffers and Lehiste 1979:150)

Based on the thresholds of Levenshtein distance 0.06 and 0.02, all the pronunciation-pairs corresponding to loans can be compared with the manual classification (Fig. 1, part B, Supplementary materials Tab. S2). We found the automatic detection of the loans (see also Van der Ark 2008 for an earlier approach) to be proportional to the estimates of the classification in both directions – that is, Indo-Iranian words into Turkic languages and vice versa) – even though at least about 30% of expert-identified loans escape the automatic detection. This proportionality is an important result, as it shows Levenshtein distance computation is not biased in one linguistic group with respect to the other one, in fact the same 50% error-rate (under-detection) is found in both.

3.4 Linguistic contact and loans

We mentioned that the linguistic research presented in this article is preliminary to an upcoming inquiry about the correlation between the cultural and genetic diversity of the very same populations. This is why we try to estimate social/cultural contact from word borrowing. According to the estimates

of PM, the percentages of borrowing (from one linguistic group into the other) are reported in table S1 (Supplementary Materials) and visualized in figure 1, part B.

The first result about the loans concerns the higher percentage of borrowing in locations close to the borders of linguistic groups (Ristan, Kaptarhona, Soj Mahalla and Orday) or where two linguistic communities live in the same place (Zarmanak and Novmetan). In all these villages the linguistic exchange seems symmetrical, except in the village of Urtoqquisloq, which, being a Turkic (Uzbek) linguistic isolate in an Indo-Iranian-speaking area, borrows more from Tajik than vice versa.

We also note that Karakalpak speakers (in the villages of Hitoj, Kokdaria, Seghe and Halqabad) seem to borrow more from Indo-Iranian than do the two Kazakh villages of Raushan and Gazli. Actually, the speakers of Gazli[7] (but not those of Raushan) come from a group recently emigrated from Kazakhstan, a country that has no Indo-Iranian speakers nearby. As expected, the Tajik speakers living in Tajikstan (Sink, Urmetan, Navdi, Nimic, Nusor) and the Kyrgyz speakers living in Kyrgyzstan (Kulanak, Ak-Muz, Tamga; Barskoon) show a very low degree of borrowing (actually the few loans are historical ones as, currently, there are no allophone neighbors from which the borrowing could have happened in recent times). Some possible reasons will be discussed in the next section.

4. Discussion

The purpose of this paper was twofold. First we provided a survey of linguistic relations in a complex area in Central Asia; we described how we designed a linguistic survey, how we computed linguistic distances between and within the two groups of languages we studied (Indo-Iranian and Turkic), and how well these reflected traditional designations. We used edit distance for this purpose, which predictably worked well. Second, we estimated the proportion of loans from a language family into the other, and we tested whether loans words might be detected automatically. This worked less well, but the automatic procedure might free larger scale studies from needing to check all candidate loan words by hand.

[7] Gazli, 90 Km from Bukhara was founded in 1958 in the middle of the Kyzyl-Kum desert to exploit natural gas resources in the region

The linguistic classifications we derived from the data are not intended to be an assessment of the historical relatedness of the linguistic varieties under study; in fact our approach—in many aspects—is more similar to socio-linguistic inquiry than to historical linguistics methodology.

With respect to our longer-term goals of understanding the parallels between genetic and linguistic diversity, we note that population genetics initially shifted from a quest for a systematic correlation to a denial of any reciprocal influence, where any correlation was seen as a by-product of the decreasing chance of human interaction when geographic distance rises (see contribution of P. Chareille in Darlu et al. 2012 and Boattini et al. 2012 for examples about migrations in recent historical times). However, a more pragmatic approach has also emerged, that is to check whether a correlation exists between the two, and to seek explanations for correlations that do emerge in the realm of cultural and demographic interaction. This is particularly the case in Central Asia where the (semi) nomadic lifestyle of many populations makes its history complex to understand.

Historical linguists have often been reluctant to provide a genealogical tree of languages, even at a macrofamily or family level, making the statistical correlation of linguistic with genetic data (available as numbers) an unapproachable issue until the quite recent spread of reliable computational linguistics methods allowing fast, reliable comparison (for an early review see Forster and Renfrew 2006). Computational approaches based cladistic methods or on network analysis, can be applied to historical linguistics with aims that are similar to glottochronology and lexicostatistics. In a similar vein, the Levenshtein distance has been developed by dialectologists as a measure of pronunciation difference enabling a kind of inference that is more geographical or social than historical. In the bootstrap tree (Fig. 2) only the first separation between Indo-Iranian and Turkic language has an historical explanation. Following clusters rely more on geographical factors than on the historical split of languages.

4.1 Networks versus linguistic distances

Network analyses make possible a new kind inquiry into linguistic phylogeny by also displaying conflicting signals that weaken the vertical evolutionary signals. A potential cause of error are faulty cognacy judgments (for instance possible chance-similarities), which increases the measured similarity

between languages and leads to an underestimation of the divergence times. Unrecognized borrowing between closely related languages would have a similar effect. Conversely, unrecognized borrowing between distantly related languages will incorrectly *depress* (not *inflate* as stated in Gray and Atkinson 2003) branch lengths at the origin of the tree and, therefore, increase the estimates about divergence-times.

We turned to linguistic distances because we wanted to geographically portray linguistic differences, regardless of their origin in time. While phylogenetic trees are better fit to provide historically reliable family trees of languages and the dates of language splits, the genetic distances we will compare (in future work) to our linguistic distances are much "noisier" because they also include the effect of migrations and other areal factors. The closeness and relative geographic contiguity of the populations we studied involve cross-migration and admixture, meaning that their phylogeny is multi-faceted and difficult to disentangle. This is why we chose to measure linguistic differentiation by the Levenshtein distance, as a way to address population differences without exclusively relying on traditional language phylogeny.

The Levenshtein method was used to measure linguistic differences between similar varieties, just as it has been applied to analyze the relations among dialects (Nerbonne and Heeringa 2010) and closely related languages (Alewijnse et al. 2007). When the linguistic differences are too great, the Levenshtein method may reach a ceiling so that it no longer reflects common provenance (Greenhill 2011), but Jäger (2013) calls this into question. While the dissimilarity of Indo-Iranian and Turkic language groups is too high to be appropriately measured using edit distance, the range of difference within each group is comparable to the differences we find among dialects in some language areas. This is why we do not discuss inter-group distances.

4.2 Effect of loans on linguistic distances

The distance matrices we computed take into account linguistic contact. Loans are not excluded from the analyses, meaning that the realization pairs corresponding to loans correspond to very low, or null, Levenshtein distances. These low edit-distances decrease the aggregated distance that is obtained from the sum of pairwise distance corresponding to each realization pair. To be sure that all loans are

recognized as null distances, we have verified their status based on the judgment of an expert. The two estimates are proportional but about 30-50% of the loans escape automatic detection. As for the reasons of the discrepancy, the most probable one is the existence of some structural disparities at work, such as for instance different phoneme inventories. Where an experienced linguist would easily see the relation between two realizations where a vowel has shifted, the Levenshtein algorithm does not. Also for this reason, all the computed linguistic distances between language-groups are overestimated, which is not a terrible concern, as we said that we do not expect the Levenshtein method to measure the actual linguistic distance between the two families perfectly. It merely needs to correlate well with the "real distances" among the varieties, and it does (Heeringa et al 2006). Put differently, and because there are no detectable borrowings within a language family, the measurements of the linguistic distances among Indo-Iranian and Turkic varieties are not biased. The lower performance of the automatic detection, though proportional, convinced us to use the loan estimates provided by the expert as a better proxy to language and population contact (Fig. 1, Tab S2 available at journal site as supplementary material).

4.3 Swadesh word list

As the reconstruction of an historical phylogeny of languages was outside our aims, the use of the Swadesh word list might be questioned. In fact, the list was designed to better assess the history of languages by including concepts that are less likely to be borrowed, thus maximizing the number of cognate pairs and, as a consequence, limiting the possibility of detecting linguistic (population) contact. However, we turned to the Swadesh list because it is of widespread use.

In this perspective, an interesting point concerns the supposed stronger historical signal conveyed by the shorter Swadesh list (100 words) when compared to its extended version (200 words), because the concepts of the first are believed to be borrowed less (Kessler 2001, McMahon and McMahon 2005). This seems to be the case with the Yagnobi speakers that are more distant from the Tajik group in the MDS plot (not shown) based on the shorter Swadesh list than in the MDS plot based on the longer list (Fig. 3, part B). This phenomenon matches fieldwork observations, where we noticed that the Yagnobi varieties spoken in Dughova and Safedorak are lexically close to Tajik.

While the progressive replacement of the original Yagnobi vocabulary is related to the endangered status of this language (12,000 speakers in 2004 according to the Ethnologue 2015) and to the resettlement of this people to Zafarabad in 1970s,[8] the concepts described in the Swadesh list have resisted replacement, in particular those of the short version (the Tajik/Yagnobi separation Fig. 2 is supported by a bootstrap score of 100%). Of course, the distances assayed by the two lists correlate nearly perfectly, as we noted in the results section.

4.4 Sociolinguistic aspects

4.4.1 Homogeneous or areally unstructured lexical diversity. We encounter homogeneity in Kyrgyzstan (Kulanak, Ak-Muz, Tamga, Barksoom) where all the different speakers used almost exactly the same words for the Swadesh concepts. This level of homogeneity, in villages that can be quite distant (a full day by car), may be the result of school education because our informants went to school during the times of the USSR, when (secondary school) instruction mainly took place in Russian.[9] Nevertheless, even if another normalization process (loss of diversity after the collapse of the USSR and the rise of national linguistic policies) were an explanation, we would expect the same phenomenon to arise among the Tajik speakers from Tajikstan (villages of Sink, Urmetan, Navdi, Nimich and Nusor) that are located at geographic distances that are comparable to those existing between the Kyrgyz sites. Actually the Tajik speakers from Tajikstan are less homogeneous and linguistically more distant from one another than the comparable Kyrghyz sites. We found that this greater variability is not areally structured, because the boostrap analysis shows no subgroups within the Tajik cluster of Fig. 2 (clusters T1 and T2 correspond to the speakers outside the country). The reason for this lack of areal structure is not obvious, and at the moment we are unable to explain it. Concerning the homogeneity of Kyrghiz varieties, it could also be that their semi-nomadic lifestyle[10] (not shared by the Tajiks) enabled long-range contacts among distant groups, thus retarding the lexical

[8] Zafarabad is located in the northern Tajikistan plain, while the homeland is the Yagnob Valley, north-west Tajikistan, between the southern slope of the Zarafshan Range and the northern slope of the Gissar Range.
[9] Our informants went to school at the times of the USSR, when Russian was the official language in schools. Nevertheless, some school teaching in Kyrghyz was tolerated in remote areas like many of those we sampled (Derbisheva 2009).
[10] They were forced to settle as recently as the Soviets' time.

divergence that customary traditional tribal meetings have further hampered. Of course, the Tajiks and Kyrgyz speakers living where the official language is the same are those showing the lowest rate of borrowing from the other language family, respectively Turkic and Indo-Iranian. We will return to this.

Finally, we note that the linguistic diversity we analyzed is lower than the one existing in the region in general, because linguistic interviews were conducted in a very formal context that is very far from natural language conditions. Conversely, recorded varieties are probably quite conservative because we enrolled, in large majority, middle-age male informants that, in general, have been found to be more conservative than females (Labov 1990; Chambers 1995: 102-103).

4.4.2 Linguistic isolation and contact

In our sampling design, chosen by the colleagues involved in the genetics part of the project, there are several speakers of the different languages that live in a country whose official language is different. Among them, there are the Kyrgyz speakers of Orday (Ubzbekistan), the Tajik speakers of Ristan, Kaptarhona, Ohalik, Kamangaron (Uzbekistan) and finally the Uzbek speakers of Urtoqquisloq (Tajikstan). In all the cases (besides the single informant from Ohalik), the speakers of the villages we mentioned are clustered together with those of the "motherland" though they systematically belong to *specific* clusters in the bootstrap tree of figure 2 (Orday → cluster Ki2; Ristan → cluster T1; Kaptarhona → cluster T2; Kamangaron → cluster T1, Urtoqqisloq → cluster Uz1). These speakers are somewhat "isolated" because they live in a country whose official language is part of a different language family (the Indo-Iranian speakers of Ristan-Kaptarhona-Kamangaron live in the Turkic-speaking Uzbekistan, and the Turkic speakers from Urtoqqisloq live in the Indo-Iranian-speaking Tajikstan), with the only exception of the Kyrgyz speakers living in Orday that is located just across the Uzbekistan border (Kyrgyz and Uzbek are both Turkic languages). The higher borrowing is easy to explain because all these linguistic *pseudo*-isolates (we say 'pseudo' because there are other Tajiks in Uzbekistan and vice versa) are constantly exposed to an official language that is quite different from the language spoken at home. In this context, their belonging to specific groups of the bootstrap tree (Fig. 2) can be interpreted in two different ways. One explanation is that some borrowed words,

specific to these communities living "abroad", may decrease the overall linguistic distances among them and, conversely, inflate those with the Tajiks living in Tajikstan.[11] Another explanation is that Tajik speakers living outside Tajikstan have maintained a vocabulary that has been less conditioned by the normalization process we addressed in the previous section. Finally, the consensus dendrogram suppresses internal structure that is not present reliably, meaning that the Tadjik varieties higher in the node may appear less similar because their clustering algorithm experience contention about the internal nodes.

As we noted in the body of the text, it is clear that loan words from language *l* in *l′* do not show that the *l′* speakers borrowed the words directly from *l*. It is always possible that a third party or third parties were involved. We emphasize therefore that loan words are to be interpreted as evidence of direct or indirect contact, perhaps via third parties.

4.4.3 Bilingualism

As we have seen, in Central Asia it is commonplace to find ethnic groups speaking a language different from the official one, and one similarly often finds bilingual ethnic groups, such as those in Zarmanak and Novmetan, where our informers could speak Uzbek and Tajik perfectly, although one language was preferred at home. Their bilingualism, together with phenomena described in the preceding section, is another reason for the high number of words borrowed from the other language family (Fig. 1, part B).

4.4.4 Kazakh and Karakalpak speakers

Kazakh and Karakalpak speakers deserve a separate discussion because the two languages seem really close, at least as far as our samples showed. The Kazakh speakers from Raushan (cluster Ka1 in Fig. 2) are clustered together with the Karakalpak ones from Shege-Halqabad (Kk1, in Fig.2), the ones

[11] Borrowings from Uzbek seem to be similar for many of the Tajik speakers living in Uzbekistan. For example the word for 'forest' (urmon) is used in 5 of 6 sites. Less frequent is the use of the Uzbek words for "sand' (qum), 'seed' (urok/uruk), 'to hunt' (aw), 'to think' (ujla-), 'to turn' (ajnali), 'to squeeze' (qisi), 'lake' (kul), and 'cloud' (bulut). In Zarmanak and Novmetan we noted the widespread of the 'sea' (deŋiz), 'mountain' (toγ), 'father'(ota), 'mother'(ona). We have not attempted to quantify this effect and compare to the effect of borrowing in other areas.

from Kokdaria (Kk2 and Kk3 in Fig.2) and the self-defined Uzbeks from Hitoj (they actually speak Karakalpak → cluster Kk^{UZ} in Fig. 2). This finding is in agreement with Kirchner (1998) who treats Karakalpak as "so closely related" (p.318) that he describes only the aspects that differ from Kazakh. (See also Russian-lg ref from review if we can get it.) In a different way, the Kazakh speakers from Gazli belong to a different group (Ka2 in Fig. 2), which corresponds to a group of Kazak workers that migrated to Gazli (Uzbekistan) 470 kilometers (as the crow flies) far from Raushan (Uzbekistan). This discrepancy cannot be explained as a misclassification of the Kazakhs from Raushan because every speaker enrolled in the study was questioned about his or her ethnical affiliation, according to recommendations of a competent ethnologist. The reason why the two non-Karakalpak ethnic groups we sampled in Karakalpakstan (the Uzbeks from from Hitoj and the Kazakhs from Raushan) actually speak Karakalpak is not clear to us. Either they were originally Karakalpaks that later embraced another ethnical identity (for example for social reasons such as in order to acquire prestige, and thereby obtain access to certain jobs) or, indeed, they belong to a different ethnic group that has completely lost its language. This question is however interesting because it is an exception to the expected undividable transmission unity of the "cultural package" (traditions, beliefs, language). What can be outlined is that the inhabited part of Karakalpakstan is quite small and surrounded by the desert, corresponding to a quite isolated region where culture assimilation may happen in a way that is different from regions that are less isolated and more extended. Finally, the fact that all the speakers located in Karakalpakstan, whatever their ethnic affiliation, use a percentage of words borrowed from Indo-Iranian that is higher than average cannot be explained as an increased contact with Tajik speakers but ought to be seen as a secondary effect of the Ubzek language they are exposed to. The following Karakalpak words are Indo-Iranian in origin, and are also found in Uzbek: Karakalpak: [kalta/kʲeltʲe], Uzbek [kalta/qalta/kɛlte] 'short'; Karakalpak [tʲerek/ tʲerɛk/daraxt], Uzbek [tʲerɛk/daraxt] 'tree'; Karakalpak [gʉl/gʊl], Uzbek [gʉl/gʊl/gyl] 'flower'; Karakalpak [gʉʃ/gʊʃ/gøʃ], Uzbek [gʉʃ/gøʃ/gʉʃt /gwʉʃt/gʊʃ/], 'meat'; see, too, the pronunciations in S2 for the concepts 'fruit', 'seed', 'egg', 'horn', 'tail', 'feather', 'river','dust', 'old', and 'left' . While we find no Indo-Iranian words in Karakalpak that are not also attested in Uzbek, we do occasionally find Indo-Iranian loans in

Uzbek that have not moved on into Karakalpak (see S2, concepts ‘dig’, ‘say’, ‘sing’, ‘fire’), confirming that the path led from Indo-Iranian through Uzbek into Karakalpak.

In reality, according to the Uzbek norm, many words that are close to Tajik have been deliberately replaced, for political purposes, by others that do not correspond to the spoken language. This is similar to what happened to American English, which has diverged from British English as a result of deliberate intervention to reform spelling, and as the result of cultural independence when new words were adoped for concepts that also existed in the United Kingdom.

4.5 Perspectives of investigation

As we mentioned already, our next scientific endeavor will be to compare the patterns of genetic variability with those of the linguistic differentiation described in this paper. As a working hypothesis, we expect the groups that are bilingual and/or that use many loan words to be more admixed genetically and vice versa. A good mapping of the ethnic groups would also be useful to see which languages are in direct contact or not. The only comprehensive documentation available at the moment (CIA 1993) is not fully convincing, because many ethnic groups are located outside the country to which they culturally “belong” (Tajiks in Uzbekistan, Uzbeks in Tajikistan or Kyrgyzstan), all of whom appear in the documentation as less numerous than our fieldwork revealed. It goes without saying that we could not approach the authors of the map to question them about the methodology used to obtain it. This is why we invite the reader to consult such a map to get a rough idea of the capricious human geography of the region and to appreciate the extent of uninhabited territories.

Concerning Karakalpakstan, it will be interesting to see whether the Uzbek and Kazakh groups that (probably) lost their language exhibit a genetic difference from the Karakalpaks than exceeds the average difference among the Karakalpaks group itself. As far as cultural anthropology is concerned, Karakalpakstan is an exceptionally interesting area deserving further research.

Linguistically, our paper suggests that there is a place for further work in the automatic detection of loan words. We used a very rough measure of pronunciation difference, and we would expect the detection rate to improve if we employed a more sensitive measure, but we leave this, too, to future work.

References

Alewjinse B., Nerbonne J., Van der Veen L. Manni F. 2007. A Computational Analysis of Gabon Varieties In Petya Osenova et al. (ed.) Proceedings of the RANLP Workshop on Computational Phonology Workshop at the conference Recent Advances in Natural Language Phonology Borovetz (Bulgaria). pp.3-12.

Andreev N.D, Sunik O.P. 1982. O probleme rodstva altajskix jazykov i metodax ee rešenija, *Voprosy jazykoznanija*, 2 : 26-35.

Balci B., Ibraguimov Kh., Mansourov Ou. et Uhrès J. (2001). *Dictionnaire ouzbek-français*. L'Asiathèque, 325 pp.

Baskakov N.A. (Ed.). 1966. *Tjurkskie jazyki, Jazyki narodov USSR II*, Moscow, USSR: Nauka, 530 pp.

Boattini A., Lisa A., Fiorani O., Zei G., Pettener D., Manni F. 2012. General Method to Unravel Ancient Population Structures Through Surnames. Final Validation on Italian Data. *Human Biology* 84(3): 235-270.

Chaix R, Austerlitz F, Khegay T, Jacquesson S, Hammer MF, Heyer E, and Quintana-Murci. 2004. The genetic or mythical ancestry of descent groups: lessons from the Y chromosome. *American journal of human genetics* 75:1113-1116.

Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, Austerlitz F, and Heyer E. 2007. From social to genetic structures in central Asia. *Current biology*, 17:43-48.

Chambers J. 1995. Sociolinguistic theory. Linguistic variation and its social significance. Oxford, UK and Cambridge, USA: Blackwell Publishers.

CIA (Central Intelligence Agency), 1993. Major ethnic groups in Central Asia, map n°729792 9-93 [25 x 34 cm, color]. CIA, Washington DC (USA). Accessed through the website of the Library of Congress of the USA (catalog number 93686639): www.loc.gov/item/93686639.

Darlu P, Bloothooft G, Boattini A, Brouwer L, Brouwer M, Brunet G, Chareille P, Cheshire J, Coates R, Dräger K, Desjardins B, Hanks P, Longley P, Mandemakers K, Mateos P, Pettener D, Useli

A, Manni F. 2012. The family name as socio-cultural feature and genetic metaphor: from concepts to methods. *Human Biology* 84(2):169-214.

Derbisheva Z. 2009. Jazykovaja politika i jazykovaja situacija v Kyrgyzstane. *Russian Language Journal*, 59: 1.

Ethnologue 2015. Ethnologue, languages of the world. Summer Institute of Linguistics (SIL) International Publications, Dallas (TX), USA. Online version at http://www.ethnologue.com

Forster P. and Renfrew C. (Eds.). 2006. Phylogenetic methods and the prehistory of languages. McDonald Institute for Archaeological Reserch, Cambridge (UK).

Gooskens C, Heeringa W, & Beijering K. 2008. Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing*, *2*(1-2), 63-81.

Gooskens C., & Heeringa W. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, 16(3):189-207.

Gray R.D. and Atkinson Q.D. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature. 426(6965):435-9.

Greenhill S. J. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, *37*(4): 689-698.

Heeringa W. Measuring Dialect Pronunciation Differences using Levenshtein Distance. PhD thesis. University of Groningen, 2004.

Heeringa W., Nerbonne J. and Kleiweg P. 2002. Validating dialect comparison methods. In W. Gaul and G. Ritter (eds.) *Classification, automation, and new media*. *Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation.* 445-452. Berlin: Springer.

Heeringa W., Kleiweg P., Gooskens C., Nerbonne J. 2006. Evaluation of String Distance Algorithms for Dialectology. In: J.Nerbonne & E.Hinrichs (eds.) Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006. 51-62.

Heyer E, Balaresque P, Jobling MA, Quintana-Murci L, Chaix R, Segurel L, Aldashev A, and Hegay T. 2009. Genetic diversity and the emergence of ethnic groups in Central Asia. *BMC Genetics*, 10:49.

Heyer E, Brazier L, Segurel L, Hegay T, Austerlitz F, Quintana-Murci L, Georges M, Pasquet P, and Veuille M. 2011. Lactase persistence in central Asia: phenotype, genotype, and evolution. *Human Biology* 83:379-392.

Holman, E. W., Wichmann S., Brown C. H., Velupillai V., Müller A. and Bakker D. 2008. Explorations in automated language classification. *Folia Linguistica*, 42(3-4): 331-354.

Jacquesson F. 2002. Les parlers karakalpak dans leur contexte. *Cahiers d'Asie Centrale*, "Karakalpaks et autres gens de l'Aral, entre rivages et deserts" 10: 93-137. Tachkent, Aix en Provence, Uzbekistan/France: Edisud.

Jacquesson Sv. 2002. Parcours ethnographiques dans l'histoire des deltas. *Cahiers d'Asie Centrale*, 10: 51-92. Tachkent, Aix en Provence, Uzbekistan/France: Edisud.

Jäger G. 2013 Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3(2): 245-291.

Jeffers R. J. and Lehiste I. 1979. *Principles and methods for historical linguistics*. Cambridge: MIT Press.

Johanson, Lars and Csató, Éva Á. (eds.), The Turkic languages, 333-343. London, UK: Routledge.

Ju J. 2002. Combined Algorithms for Constrained Estimation of Finite Mixture Distributions with Grouped and Conditional Data. MA Thesis, McMaster University, Ontario, Canada.

Junusaliev B. M. 1966. Kirgizskij jazyk. In : V.V. Vinogradov, (Ed.), *Jazyki narodov USSR , 2, Tjurkskie jazyki.* Moscow, USSR: Nauka.

Junusaliev B. M. 1966. « Kirgizskij jazyk », *in* : V.V. Vinogradov, red., *Jazyki narodov USSR* , 2, *Tjurkskie jazyki,* Nauka : 482-505.

Kassian, A. 2015. Towards a formal genealogical classification of the Lezgian languages (North Caucasus): Testing various phylogenetic methods on lexical data. *PloS one, 10*(2), e0116950.

Kerimova A.A. 1959. *Govor tadžikov Buxary, Izd. vostočnoj literatury,* Moscow, USSR . 163 pp.

Kessler B. 2001. The significance of word lists. CSLI Press, Stanford (USA).

Kirchner M. 1998. Kazakh and Karakalpak. In: L. Johanson and É. Csató (Eds.). The Turkic languages, Taylor & Francis. pp. 318-332.

Labov W. 1990. The intersection of sex and social class in the course of linguistic change. *Language variation and change* 2: 205-254.

Levenshtein V. I.. Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory, 10:707–710, 1966.

List J. M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

Manni F. 2010. Sprachraum and genetics. Dans : A. Mameli, R. Kehrein, S. Rabanus (eds.) *Mapping language*. Mouton de Gruyter, Berlin-New York. (pp. 524-541).

Manning, C. D., & Schütze, H. 1999. *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Martinez-Cruz B, Vitalis R, Segurel L, Austerlitz F, Georges M, Thery S, Quintana-Murci L, Hegay T, Aldashev A, Nasyrova F, and Heyer E. 2011. In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *European Journal of Human Genetics*, 19: 216-223.

McMahon, A and McMahon R. 2005. Language classification by numbers. Oxford University Press, Oxford (UK).

Menges K.H. (1968). *The Turkic languages and peoples – An Introduction to Turkic studies*, Ural-Altaische Bibliothek, Otto Harrassowitz, Wiesbaden : 163 p.

Montemagni, S., Wieling, M. de Jonge, B and Nerbonne, J. (2013). Synchronic Patterns of Tuscan Phonetic Variation and Diachronic Change: Evidence from a Dialectometric Study. *LLC: Journal of Digital Scholarship in the Humanities* 28(1): 157-172.

Moukhtor Ch., Ibraguimov Kh. and Mansourov Ou. 2003. *Dictionnaire Tajik-français*, published by "Langues & Mondes-L'Asiathèque et IFEAC", Paris, France, 357 pp.

Nerbonne, J. (2009). Data-Driven Dialectology. Language and Linguistics Compass, 3(1): 75-198.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., & Leinonen, T. (2011). Gabmap — a web application for dialectology. *Dialectologia: revista electrònica*, 65-89.

Nerbonne J, Heeringa ., & Kleiweg P. 1999. Edit distance and dialect proximity. In: D SAnkoff & J Kruskal (eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed., v-xv*. Stanford: CSLI Press.

Nerbonne J., Kleiweg P., Heeringa W., Manni F.. 2008. Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker (eds.) Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society, Berlin: Springer. 647-654. (Studies in Classification, Data Analysis, and Knowledge Organization)

Nerbonne J., & Siedle C. 2005. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72(2): 129-147.

Prokić J., and Nerbonne J. 2008. Recognising groups among dialects. *International Journal of Humanities and Arts Computing*, 2(1-2) :153-172.

Prokić, J., Nerbonne, J., Zhobov, V., Osenova, P., Simov, K., Zastrow, T., & Hinrichs, E. (2009). The computational analysis of Bulgarian dialect pronunciation. *Serdica Journal of Computing*, *3*(3): 269-298.

Rastogueva V.S. 1963. *Očerki po tadžikskoj dialektologii, 5 : Tadžiksko-russkij dialektnyj slovar',* AN USSR , Moscow, 250 pp.

Rastorgueva V.S. 1964. *Opyt sravnitel'nogo izučenija tadžikskix govorov,* Nauka, Moscow, USSR , 188 pp.

Segurel L, Austerlitz F, Toupance B, Gautier M, Kelley JL, Pasquet P, Lonjou C, Georges M, Voisin S, Cruaud C, Hegay T, Aldashev A, Vitalis R, and Heyer E. 2013. Positive selection of protective variants for type 2 diabetes from the Neolithic onward: a case study in Central Asia. *European Journal of Human Genetics,* 21:1146-1151.

Segurel L, Lafosse S, Heyer E, and Vitalis R. 2010. Frequency of the AGT Pro11Leu polymorphism in humans: Does diet matter? *Annals of Human Genetics* 74:57-64.

Segurel L, Martinez-Cruz B, Quintana-Murci L, Balaresque P, Georges M, Hegay T, Aldashev A, Nasyrova F, Jobling MA, Heyer E and Vitalis R. 2008. Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. *PLoS Genetics* 4:e1000200.

Šimičić, L., Houtzagers, P., Sujoldžić, A. and Nerbonne, J. 2013. Diatopic Patterning of Croatian Varieties in the Adriatic Region. *Journal of Slavic Linguistics*, 21(2): 259-301.

Sturrock K. and Rocha J. 2000. A multidimensional scaling stress evaluation table. *Field methods*, 12: 49-60.

Swadesh M. 1955. Towards greater accuracy in lexicostatistic dating, *International Journal of American Linguistics*, 21: 121-137.

Swadesh M. 1972. "What is glottochronology". In: Swadesh M (Ed.) *The origin and diversification of languages*, London, UK: Routledge and Kegan Paul, pp. 271-284.

Van der Ark 2008. Comparing languages and dialects in Central Asia. M.A. Thesis. University of Groningen.

Van Der Ark R., Mennecier P., Nerbonne J., and Manni F. 2007. Preliminary Identification of Language Groups and Loan Words in Central Asia. In *Proceedings of the RANLP Workshop on Computational Phonology* RANLP : Borovetz. pp. 12-20.

Wells, J. 1997. SAMPA computer readable phonetic alphabet. In Gibbon, D., Moore, R. K., & Winski, R. (Eds.). *Handbook of standards and resources for spoken language systems*. Walter de Gruyter, Berlin. Appendix B.

Wichmann S. at al. 2013. The ASJP Database (version 16). Avail. at http://asjp.clld.org/

Wieling M., Margaretha E., & Nerbonne J. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, *40*(2): 307-314.

Wieling M., and Nerbonne J. 2015. Advances in dialectometry. *Annual Review of Linguistics.* 1(1): 243-264.Wieling M., Shackleton R. Jr. and Nerbonne J. 2013. Analyzing phonetic variation in the traditional English dialects: Simultaneously clustering dialect and phonetic features. *LLC: Journal of Digital Scholarship in the Humanities* 28(1) :31-41.