Historical linguistics and molecular anthropology

Brigitte Pakendorf

The Routledge Handbook of Historical Linguistics

Edited by
Claire Bowern and Bethwyn Evans

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

29

Historical linguistics and molecular anthropology

Brigitte Pakendorf

1 Introduction[1]

At first glance, it might come as a surprise to find a chapter on molecular anthropology in a handbook of historical linguistics. And yet, as will be outlined below, molecular anthropological studies can provide insights into prehistoric processes that may have had an impact on language change, thus offering the potential of deepening our understanding of such changes. The reasons for this potential are that both 'genes' (DNA molecules) and languages are passed on by human beings through social interactions, and both genes and languages can retain traces of prehistory, leading to the expectation that genes and languages should coevolve. As will be outlined briefly in section 2 below, this potential coevolution of genes and languages has stimulated research predominantly among geneticists who are interested in elucidating whether cultural factors like language might have an impact on biological evolution. A different approach to genetic insights into language change, which is driven by questions concerning language evolution (specifically, contact-induced language change) rather than genetic evolution, is at the heart of this chapter and will be described in section 3. Since this is still a very young field of research, the focus will be not so much on a review of results, but rather on introducing this interdisciplinary approach to population and language contact and the insights it can provide into processes underlying language change. For readers who might need a (re-)introduction to genetics, the Appendix provides a brief overview of some of the most important concepts needed to follow this chapter.

Both approaches to combining genetic and linguistic data, the 'coevolution' and the 'contact' approach, are based on certain parallels between genes and languages: first of all, most people grow up speaking the language of their parents as their first and dominant language, so that in this case both genes and languages are passed on from parents to offspring. Second, when people migrate, they carry both their genes and their languages with them to their new place of settlement – as illustrated by peoples of European ancestry speaking European languages now residing in the Americas, in South Africa, and in Australia and New Zealand. Furthermore, both genes and languages may retain traces of contact – genetic admixture (called 'gene flow') and contact-induced changes such as loanwords, calques, or structural changes, respectively. However, there are also significant differences between genes and languages that need to be taken into account when trying to compare them: first of all, while genes are always transmitted only in the vertical line, from parents to offspring, the same does not hold for languages, which can be and often are transmitted horizontally

through contact with the local community (cf. Stanford, this volume). A further significant difference concerns the different time depth of evolution of genes and languages: languages change and differentiate much faster than genetic material does, so that the time depth for linguistic reconstruction and that for genetic reconstruction differ by at least one order of magnitude. Thus, it is frequently not possible to use the genetic relationship of populations as an indication of the genealogical relationship of the languages they speak (contrary to widespread belief): if two populations share a common ancestor 30,000 years ago, that ancestor might have spoken a language ancestral to the languages now spoken by its descendants, but this need not be the case, and with current linguistic methodology it can hardly be established. Nevertheless, this difference in time depth does not absolutely preclude fruitful comparison of linguistic and molecular anthropological data, since the latter can also retain traces of relatively recent events, such as migrations dating back only a few hundred years, and thus provide direct insights into contact-induced changes (cf. section 3).

2 Coevolution of genes and languages

As mentioned above, the parallel evolution of genes and languages has stimulated research predominantly among geneticists. Two major models can be distinguished in this line of research: the 'branching' model, which assumes that languages and genes evolve through successive splits followed by isolation, and the 'isolation-by-distance' model, which assumes that languages and genes diverge gradually over geographic distance, with contact between geographically close entities leading to the exchange of linguistic features and/or genetic material, while such contact diminishes as spatial distance increases. In the perspective of the branching model, a lack of correlation between linguistic and genetic distances is interpreted as being due to replacement of either languages or genes, while in the isolation-by-distance model boundaries of abrupt genetic change that coincide with linguistic boundaries are taken as an indication that the latter represent barriers to gene flow – i.e. that peoples speaking unrelated languages tend not to intermarry as frequently as peoples speaking related languages (Barbujani 1991: 151–152; McMahon 2004: 3). On a global scale, there do indeed appear to be statistically significant positive correlations between genes and languages that have been interpreted as supporting the branching model of language–gene coevolution (Barbujani 1991: 152–153) – but it remains an open question what the biological and linguistic significance of such correlations might be (cf. McMahon 2004: 6).

Most of the studies investigating language–gene coevolution are based on correlations between genetic and linguistic distances. While there are several methods to estimate genetic distances between populations that are tailored to the type of genetic data involved, estimating linguistic distances is frequently a major problem. To obtain accurate distance measures comparable to those employed for the genetic data, the distance measure should be based on actual linguistic data; however, this requires a good prior knowledge of the languages involved, is time-consuming, and cannot be undertaken without linguistic expertise (cf. Dunn, this volume). Therefore, several studies have taken a short-cut approach and assigned distances according to the position of the linguistic entities in some genealogy – with dialects of the same language being assigned the smallest distance, and languages belonging to different phyla the largest (e.g. Poloni *et al.* 1997; Wood *et al.* 2005; Belle and Barbujani 2007). In general, the genealogies used for this purpose are those compiled by Ruhlen (1987) – disregarding the fact that these are often highly controversial among linguists. Needless to say, this approach cannot provide any insights of value to historical linguists. Not only are the linguistic 'distances' far too crude to provide any meaningful basis for comparison, but the

resulting correlations don't provide any insights into language change. In addition, while a significant and positive correlation between genetic and linguistic distances is interpreted as supporting the coevolution of languages and genes, lack of such a correlation is often argued away with post hoc explanations (McMahon 2004: 11).

However, the investigation of language–gene coevolution is being refined, and more recently several studies have appeared that attempt to address specific historical questions by comparing genetic and fine-scaled linguistic distances based on typological features, the proportion of shared cognates, or phoneme inventory size (Lansing *et al.* 2007; Hunley *et al.* 2008; de Filippo *et al.* 2012; Hunley *et al.* 2012). These studies demonstrate that at a local scale it is isolation by distance (i.e. varying amounts of genetic and linguistic contact), rather than splits followed by isolation, that plays a significant role in shaping both genetic and especially linguistic evolution. This indicates that a refined knowledge of the processes at play in such situations of contact are important for our understanding not only of human evolution, but also of language change, as will be outlined in the following sections.

3 Molecular anthropological insights into externally induced language change

For a long time, the a priori assumption in historical linguistics was that internally motivated change is the default and contact-induced changes are rare and should be resorted to as an explanation only in exceptional cases:

> For well over a hundred years, mainstream historical linguists have concentrated heavily on system-internal motivations and mechanisms in studying language change. The methodological principles embodied in the powerful Comparative Method include an assumption that virtually all language change arises through intrasystemic causes.
>
> (Thomason and Kaufman 1988: 1)

However, it has become abundantly clear in the past 25 years that externally motivated language change is in actual fact commonplace (see also Lucas, this volume) – in good accordance with the results from the fine-scaled language–gene coevolution studies mentioned above – leading to a plethora of innovations, from simple loanwords to phonological influence and structural changes. Such external motivation can encompass a range of sociocultural contact situations: from mere 'culture contact' in which actual interactions between different speech communities remain at a minimum, to long-term societal bilingualism – especially common in small-scale societies such as those of Melanesia or Africa – to language shift, in which one speech community gives up its heritage language in favour of the language of another community, often that of the socioculturally dominant group. These different types of contact are expected to lead to different kinds of changes in the languages concerned – but the nature of the correlation is still not well-established. And yet, it would be a further useful historical tool if we knew more about which kinds of contact situation lead to which kinds of linguistic change, since this could then enable us to deduce something about historical sociocultural situations from current linguistic data (cf. Ross 2003; Ross 2013; Epps, this volume). While we might never attain diagnostic precision when it comes to identifying the nature of prehistoric contact situations from the linguistic changes observed, since too many variables can affect the outcome of language contact, we might nevertheless be able to refine our deductive ability. However, currently our endeavours to identify correlations between contact situations and their linguistic impact are hampered by the lack of historically attested contact situations with known linguistic outcomes with which to calibrate our diagnostic tools.

This is where the field of molecular anthropology can be of use. The underlying assumption in this approach is that the mismatch between the genetic and linguistic affiliation of communities can provide meaningful insights into past historic events (see Heggarty, this volume, for the historical processes that might lead to such mismatches). Since humans can learn to speak any language, and even more than one language, irrespective of their genetic background, they can give up their ancestral language in favour of a new one, leading to a mismatch between their genetic and their linguistic affinities. Similarly, groups of people can intermarry with communities of a different genetic and linguistic background, but maintain their language – this, too, leads to a detectable mismatch between their genetic and their linguistic affinities. Furthermore, thanks to two useful parts of the human genome, the maternally-inherited mtDNA and the paternally-inherited Y-chromosome, molecular anthropological investigations can highlight cases of sex differences in admixture. Conversely, they can also provide insights into cases where there has been no physical interaction, which can in turn be as informative for the investigation of different kinds of language contact as those cases in which we find evidence of intermarriage. Molecular anthropological studies are therefore beginning to play a more and more important role as a subsidiary field to enhance our knowledge about prehistoric language contact, as will be discussed in the following sections.

However, while molecular anthropological analyses permit some important insights into the prehistory of populations which can indeed be of use for studies of language contact, as will be illustrated in the rest of this section, they are of course not without problems. First of all, the results of such studies can only be as good as the data that went into the analyses, and if the data are not from properly chosen ethnolinguistic groups, the results might be hard to interpret or completely unusable for linguistic purposes. As molecular anthroplogists tend to be specialists in genetics, not linguistics or even anthropology, they do not necessarily know all the interesting and important distinctions between ethnolinguistic groups. Quite frequently, one finds only geographically defined 'populations' (e.g. 'Bantu speakers from Kenya'), or populations that are lumped (e.g. '!Xuun/Khwe'), irrespective of the fact that they might speak very different languages and may have had very different histories. In general, the best results are obtained when linguists and molecular anthropologists collaborate closely (cf. McMahon 2004: 10), as is happening more frequently (e.g. Quintana-Murci *et al.* 2008; Berniell-Lee *et al.* 2009; Barbieri *et al.* 2012; Pickrell *et al.* 2012; Barbieri *et al.* 2013a).

Second, admixture between genetically distant populations is much easier to detect than between genetically similar groups. In some cases, populations in a given geographic area might be genetically indistinguishable (this frequently holds for mtDNA, since patrilocality is more common than matrilocality), so that it might be impossible to address detailed questions of recent contact between individual groups. However, in such cases it might help to include autosomal DNA, since this has a higher resolution than the uniparental markers.

Lastly, it should be noted that obviously genetic analyses can only pick up signals of contact when that contact has resulted in offspring, and only when there were sufficient numbers of offspring to result in a detectable signal in the generally fairly small samples that are analysed. For example, if 20 individuals selected at random from a population are included in a study of mtDNA or Y-chromosome variation, the detectable level of admixture would be 5 per cent or more (one individual out of the 20 in the sample). In addition, it should be kept in mind that male-biased admixture does not necessarily involve close social interactions – men can easily father children without playing any role in their upbringing. Thus, one cannot necessarily deduce that all male admixture would have involved social interactions – and since linguistic changes, especially those of a structural nature, require close interactions between speakers of different languages, Y-chromosomal results might be

misleading. Nevertheless, it is unlikely that high levels of Y-chromosomal admixture would be the result of only ephemeral physical contact, so that we can still use the Y-chromosome to investigate male-specific aspects of prehistory.

3.1 Correlations between linguistic and genetic contact

Before proceeding with an overview of the potential genetic and linguistic correlates of different contact situations, it is important to emphasise that the scenarios sketched here are of necessity overly simplistic, for two reasons: first, it becomes very difficult, if not impossible, to test highly complex models containing many factors where each individual factor's contribution to the outcome is hard to disentangle; it is easier to start off with only a few simple parameters to test and to refine the model based on initial results. Second, the results obtained with molecular anthropological investigations do not permit insights into the societal or individual factors at play in contact situations, such as attitudes towards foreign linguistic items or the psycholinguistic dominance of speakers. Molecular anthropology is thus only one subsidiary discipline that can aid our understanding of the processes involved in language contact; other disciplines, such as social anthropology, sociolinguistics, and psycholinguistics, should ideally be included in our study of language contact as well (cf. Ross 2013).

The most superficial type of contact is 'culture contact' (corresponding to level 1 in Thomason and Kaufmann's [1988: 74] borrowing scale); this is expected to take place in the absence of close social and physical interactions (and in the absence of bilingualism; Ross 2003: 192) and to lead merely to the borrowing of words related to novel technologies or cultural innovations. Without close social interactions, such culture contact cannot lead to notable levels of genetic admixture between the populations in contact, so that the expected genetic correlate is a lack of detectable gene flow between the populations. This can be illustrated with the Kalmyks, nomadic pastoralists speaking a Mongolian language who migrated to southern Russia 300 years ago. They show no evidence of admixture with Russians, their new neighbours, in analyses of either the maternally inherited mtDNA or the paternally inherited Y-chromosome (Nasidze *et al.* 2005), and their language also does not show any influence from Russian, other than loanwords[2] (Bläsing 2003: 230).

Of course, lack of detectable gene flow does not necessarily mean lack of language contact, since communities can be bilingual even in the absence of intimate social/physical contact – for instance when speakers of socially subordinate languages are bilingual in the language of the socially dominant group without widespread intermarriage between the communities. This can be exemplified by the Sakha (Yakuts) and Mongols of northern Asia: the Sakha language, which belongs to the Turkic language family, contains a large amount of Mongolian loanwords – at least 13 per cent, including basic vocabulary such as kinship and body part terms (Pakendorf 2007: 295–296; Pakendorf and Novgorodov 2009). Furthermore, some structural changes are also attributable to Mongolic contact, for instance the extension of the dative case to include locative functions (Pakendorf 2007: 120–141). This points to fairly intense levels of linguistic contact; in contrast, the genetic data show hardly any physical interaction between Mongols and the ancestors of the Sakha, either in the paternal line (Pakendorf *et al.* 2006) or in the maternal line (Whitten *et al.* in preparation). The Mongols were the dominant social group over vast areas of Eurasia during the period of contact with the Sakha ancestors, and it is possible that these large sociopolitical prestige differences encouraged some Sakha to become bilingual in Mongolian. Nevertheless, the overall speech community of the Sakha ancestors most probably remained dominant in

Sakha, as evidenced by the phonological adaptation of the loanwords (Pakendorf 2007: 299, 309–311).

On the other end of the contact scale we find language shift, in which entire groups of people give up their heritage language in favour of a new language. The linguistic effects of such shift are as yet not clearly determined: since language shift leads to the loss of the language originally spoken by the shifting group, which may come to identify not only linguistically but also socioculturally with the new speech community, it can be hard to detect in the absence of historical documentation (Thomason and Kaufman 1988: 111; Ross 2013). While Thomason and Kaufman (1988: 121) suggest that both phonological and syntactic changes result from shift, Ross (2013: 37) suggests that only if the shifting group were adults (i.e. if the shift took place relatively abruptly, without a preceding period of bilingualism), are phonological changes "in the absence of other significant contact effects" as well as possibly constructional calques expected. A further outcome of shift – even when the shift is preceded by a period of bilingualism – may be the "transfer of specialist vocabulary" (Ross 2013: 37).

Since language shift involves entire communities, i.e. both men and women, the genetic concomitant of such shift situations is an influx of both mtDNA and Y-chromosomal lineages from the shifting group into the group that was the target of shift. Note that this view contrasts with that of Forster and Renfrew (2011: 1391), who propose that "… a Y-chromosomal signal may be a necessary factor, but not a sufficient one, as a predictor of language," because in their view language shift is mostly brought about by the immigration of men into an area and subsequent intermarriage with resident women speaking different languages who abandon their native language. However, while it is indeed often the case that in patrilocal societies inmarrying women are required to speak the language of their husbands' communities, so that linguistic affiliation correlates with Y-chromosomal, but not mtDNA, affinities (see the description of sex-biased intermarriage in Burkina Faso below), this is not always the case: for example, a study of the patrilocal Gilaki and Mazandarani in Iran, who speak Indo-Iranian languages, demonstrated that they are genetically close to their Indo-Iranian speaking neighbours in the maternal line (as identified with the help of mtDNA analyses), but that their paternal lineages trace back to the Caucasus (Nasidze *et al.* 2006) – i.e. in this case it appears to have been the men who gave up their heritage language in favour of the language of their wives. In direct contrast to the views of Forster and Renfrew (2011), the authors of the study conclude that "[t]he concomitant replacement of mtDNA and language after the migration of a group to a new region may thus be a more general phenomenon than previously recognised, and furthermore emphasises the role of maternal transmission of language as a means of language replacement" (Nasidze *et al.* 2006: 671). To what extent either of these two conclusions is valid still requires more research; in either case, I would maintain that sex-biased gene flow of the kind discussed by Forster and Renfrew and exemplified by the study of the Gilaki and Mazandarani is not the result of language shift proper; rather, the process of shift is individual and continuous, concerning only the inmarrying spouses, while the majority of the community continues to speak the local language. Thus, such cases are more likely to be accompanied by long-term bilingualism, with potentially very different linguistic results, as discussed below.

A clear genetic signal of language shift as proposed here, namely the concomitant introgression of both mtDNA and Y-chromosomal lineages (further supported by autosomal DNA evidence) is found in the Damara, a Khoisan[3] population in Namibia leading a life of small-stock pastoralism and foraging. The Damara speak dialects of a language which is also spoken by the Nama, a group of traditional herders of cattle as well as small livestock.

Genetically, however, the Damara are quite distinct from the Nama and all other Khoisan populations. They show only 13 per cent characteristic Khoisan mtDNA lineages (Barbieri *et al.*, 2014) and 12 per cent characteristic Khoisan Y-chromosomal lineages (Wood *et al.* 2005), and can be shown to be very closely related to the Himba and Herero (Pickrell *et al.* 2012; Barbieri *et al.*, 2014), pastoralist communities of northern Namibia who speak closely related Bantu languages. This indicates that the Damara stem from the same ancestral population as the Himba and Herero, and that they adopted their Khoisan language through contact with the Nama. To what extent this language shift had an impact on the language spoken by the Damara still requires linguistic investigation; based on a lexical study of the different Nama–Damara dialects Haacke (2008: 167) concludes that "the simplistic claim that the Damara (in their entirety) adopted the Nama language is fallacious."[4] Of course, as stated above, if the language shift was preceded by a period of bilingualism, not many changes are expected in the target language other than potentially the transfer of specialist lexicon; otherwise, one would expect predominantly phonological and perhaps some syntactic changes. Thus, a survey of basic lexicon is not expected to reveal traces of shift; rather, a study of specialist vocabulary and of phonological and syntactic changes in the Damara dialects is required.

A further contact situation is long-term bilingualism, which is expected to result in lexical and grammatical calques, and which might lead to large-scale structural changes ('metatypy'; Ross 1996: 182; 2007; 2013: 37). While the linguistic outcome of such bilingualism can be highly interesting, this contact situation can unfortunately not be linearly correlated with genetic patterns, since people can and do learn many languages outside their nuclear family context, e.g. through interactions with peers and other members of the local community (Stanford, this volume). The kind of extensive bilingualism that can lead to large-scale structural changes has been shown to take place both with widespread intermarriage (e.g. in the Vaupes area in South America; Aikhenvald 2002) and without it (e.g. in Kupwar, a village in India; Gumperz and Wilson 1971). There is thus no single genetic correlate that one could postulate as a diagnostic tool for this kind of situation, as one could find evidence of large-scale gene flow, low levels of intermarriage, or none at all, depending on the circumstances.

However, as mentioned above, one genetically detectable situation of social interaction that might result in long-term bilingualism is the frequent incorporation of spouses from a different speech community (called 'sex-biased gene flow' in the genetics literature). There are several sociocultural factors that determine such sex-biased intermarriage, e.g. post-marital residence patterns (matri- and patrilocality) or sex differences in the permeability of social barriers. Thus, it is easier for women to marry up the social ladder than for men, and it has been observed that it is women, and not men, from marginalised ethnolinguistic groups who may marry into socially dominant groups, e.g. it is women from foraging communities who marry into food-producing communities. Since sex-biased gene flow is expected to lead to differences in affiliation between the maternally inherited mtDNA and the paternally inherited Y-chromosome, it is relatively easy to detect when both mtDNA and Y-chromosomal variation are studied. For example, in Burkina Faso in West Africa, populations speaking languages belonging to two very distinct branches of the Niger-Congo phylum, Mande and Gur, are in close contact. These populations are strictly patrilocal, meaning that after marriage a woman is expected to move to her husband's homestead and to speak her husband's language; her children, however, may spend long periods of time with their maternal relatives, thus also learning their mother's language (Beyer and Schreiber 2013). As the results of genetic analyses show (Barbieri *et al.* 2012), this system has led to a complete homogenisation of the maternal gene pool of the populations concerned, while the paternal gene pool is

surprisingly structured along linguistic lines, with populations speaking languages belonging to the Mande and the Gur families being genetically distinct from each other. This female-biased gene flow between populations irrespective of the language they speak has clearly been continuing for a long time, and it might help explain the complex patterns of contact-induced changes detected in the area. For instance, morphosyntactic changes in negation and copula clause constructions have affected several languages without a single source language being discernible (Beyer and Schreiber 2013).

To summarise, the combination of molecular anthropological and linguistic studies is expected to provide insights into prehistoric contact situations and the linguistic changes resulting from them, complementing insights obtained from archaeology and history (cf. Heggarty, this volume). Most easily detectable with genetic methods are cases of 'culture contact' – i.e. the absence of intimate contact and thus lack of gene flow – as well as language shift, which is expected to result in the influx of both mtDNA and Y-chromosomal lineages from the shifting population. Cases of long-term bilingualism, which might lead to striking structural changes in the languages, are more difficult to correlate with one specific genetic outcome, although sex-biased gene flow might be one possible cause of such bilingualism. In the following, I will briefly illustrate the added insights into sociocultural processes underlying language change that can be gained from molecular anthropological studies.

3.2 Illustrating the interdisciplinary approach to language contact: a case study

The case study presented here concerns the question of how click phonemes were transferred to a small group of Bantu languages in southwestern Zambia, and my summary is based on two articles: Bostoen and Sands (2012) for the linguistic side of the story, and Barbieri *et al.* (2013a) for the molecular anthropological investigation.

One of the most salient characteristics of the so-called Khoisan languages is a heavy functional load of click phonemes. Such clicks are not characteristic of the Bantu language family, except where they have been borrowed through contact. There are several Bantu languages spoken in southwestern Africa (in the corner where the borders of Angola, Zambia, Botswana, and Namibia come together) that have borrowed clicks; here, I will focus on only one of them, namely Fwe, since the molecular anthropological study contained sufficient data only from Fwe speakers. Fwe belongs to the Botatwe subgroup of the Bantu family, and it is the only language with clicks within this subgroup. Interestingly, very few of the Fwe words with click consonants can actually be traced to an extant source language, indicating that the ancestors of the Fwe may have been in contact with a community speaking a now extinct Khoisan language. This hypothesis is confirmed by genetic data that provides evidence that genetically distinct Khoisan populations were settled in Zambia in prehistoric times; the only remaining traces of these groups are divergent Khoisan mtDNA lineages retained in some Bantu-speaking peoples (Barbieri *et al.* 2013b).

The functional load of clicks in Fwe is very small: only about 2 per cent of the vocabulary contains clicks, and the language has only one phonemic click (the dental one). The click words that can be traced to a Khoisan language contain terms pertaining to the fauna and flora of the new environment. This might be an indication that the clicks entered the language when the immigrating Bantu-speakers adopted words for local phenomena from the autochthonous Khoisan population, i.e. in a process of culture contact. However, in such a scenario the ancestors of the Fwe are not expected to have been bilingual in the Khoisan language, making it unlikely that they would have borrowed words containing consonants that are very different from those found in any Bantu language without adapting them

phonologically.[5] A more likely possibility is that the click consonants entered the ancestral Fwe language as the result of substratum influence when entire Khoisan speech communities shifted to the Bantu language of the immigrant agriculturalist peoples. As mentioned above, one of the diagnostic features of language shift is the transfer of specialist vocabulary (Ross 2013: 28–29) – such as terms for local fauna and flora. Thus, the linguistic data are compatible with a scenario of language shift of Khoisan speakers to the language of the Fwe ancestors. However, it is intriguing that in Fwe clicks are found not only in probable loanwords, but also in words of Bantu origin – these phonemes must therefore have been independently propagated in Fwe by speakers of Fwe. This indicates that a more complex process than just language shift was at play.

This is the point at which molecular anthropological data become important, since they let us distinguish between prehistoric language shift of entire communities, sex-biased intermarriage, or lack of intimate social contact, especially if the peoples in contact are of genetically distinct origins. Fortunately, this is the case with respect to Bantu and Khoisan: the latter harbour some very specific lineages which can be considered diagnostic of Khoisan gene flow into non-Khoisan populations. These are the mtDNA haplogroups L0d and L0k, and the Y-chromosomal haplogroups A-M51 (also called A3b1) and B-M112 (also called B2b). As shown by the investigation of Y-chromosomal polymorphisms defining haplogroups A-M51 and B-M112, the Fwe do not have any Khoisan-specific Y-chromosomal haplogroups, nor do they show any evidence of introgression of any other divergent Y-chromosomal lineages. This indicates that no paternal lineages of Khoisan origin entered the Fwe community. In contrast, analyses of mtDNA sequences show that nearly one quarter of the Fwe maternal lineages (24.3 per cent) are of Khoisan origin. Furthermore, the Khoisan haplogroups in the Fwe are represented by four very divergent types, indicating that the interaction between the Khoisan women and the Fwe ancestors must have been relatively intense.[6] Had the contact been restricted to just one or a very few Khoisan women marrying into the Fwe community, it would have resulted in high frequencies of only one or two Khoisan mtDNA types. The molecular anthropological results thus indicate that rather than the click words in Fwe being the result of language shift of an entire Khoisan community, they resulted from the relatively frequent marriage of Khoisan women into the ancestral Fwe community. (In this case it is fairly safe to assume marriage, rather than just sexual relationships, since the resulting offspring were considered part of the Fwe community.) The adoption of clicks as a salient marker of the Fwe language and their spread beyond borrowed words of Khoisan origin to words of Bantu origin may have been a way to flag the separate ethnic identity of a community with a sizeable proportion of non-Bantu ancestry. Nevertheless, while the marriage of women of Khoisan origin must have been relatively frequent, it clearly did not reach the levels of intermarriage detected in the populations of Burkina Faso described above: the Fwe are far from being genetically indistinguishable from Khoisan populations in the maternal line. To what extent this female-biased intermarriage resulted in other contact-induced changes in Fwe still remains to be investigated.

4 Contribution to our knowledge of language change and most interesting outstanding problems

In conclusion, it is hoped that by conducting fine-scaled molecular anthropological investigations informed by questions of linguistic interest it will be possible to gain insights into prehistoric population contact which can then be correlated with the results of linguistic studies, and that this will help us to understand the processes underlying contact-induced

language change. It is becoming increasingly clear that contact has been frequent throughout human history, and that this is an important factor influencing language change, making a thorough knowledge of the processes underlying contact-induced changes very important for our overall understanding of language history. However, a problem with this endeavour is that the majority of the world's languages are spoken in small-scale societies without long written traditions and very short histories, making it very hard to come to solid generalisations about the processes of contact-induced change. As outlined above, this is where molecular anthropology can be of help, because it can uncover prehistoric demographic events that may have played a role in language change: admixture most obviously, i.e. close physical contact between speech communities as exemplified here with the data from Burkina Faso and southwestern Zambia, but also (and just as importantly) lack of admixture as exemplified with the case of the Sakha–Mongolian contact. Molecular data can also provide evidence for population size changes which might help explain particular linguistic changes or the lack thereof – for instance, whether a reduction in population size may have led to greater dependence on neighbouring groups and a concomitant greater willingness to interact with them, as suggested for Sakha–Evenki contact (Pakendorf 2007: 320).

Since this approach is still very young, there are as yet few results and there is much scope for improvement. More detailed case studies are needed for us to be able to draw firm inferences about the possible correlations between physical/genetic contact and linguistic change. To achieve this, it is imperative that linguists and molecular anthropologists collaborate closely, since only through close collaboration can the insights from both fields be fruitfully combined. Unfortunately, such joint efforts might be hampered by the severe restrictions on molecular anthropological research in various regions of the world. Thus, genetic research into population contact is currently not feasible in Australia or North America, and other areas of the world are becoming increasingly more difficult to access. Unless these restrictions are loosened, it will not be possible to fully exploit the potential inherent in collaborative research, since as mentioned above extant sample collections or published data are frequently not amenable to fine-scaled investigations of questions of linguistic concern. A further threat to this approach is language endangerment – if languages die before linguists can document them and the possible contact-induced changes they have undergone, we will never be able to fully understand language change in small-scale societies. It can therefore only be hoped that molecular anthropological investigations of small-scale populations will become possible while these populations still speak their native languages.

Appendix: tools used to study molecular anthropology

I here provide a very brief outline of the characteristics of the major markers employed in molecular anthropological investigations. Unfortunately, at the time of writing there is still no good introduction to the field; however, a recommendable textbook is expected to appear in the foreseeable future (Stoneking, to appear). In the meantime, readers who are interested in more details are referred to Appendix 1 in Pakendorf (2007).

The human genome is divided into two very distinct parts: the bulk of our genetic material is located in the cell nucleus (nuclear DNA), while a separate type of DNA, called mitochondrial DNA (mtDNA), is found in the mitochondria, the energy-producing organelles of the cell. These parts of the genome differ considerably in size, form, and mode of inheritance: nuclear DNA is present in 23 pairs of molecules called chromosomes comprising approximately three billion nucleotides in total; mtDNA is a very small (only approximately 16,500 nucleotides in length) circular molecule that is present in hundreds to thousands of

copies inside each cell. The chromosomes that constitute the nuclear genome can be divided into 22 pairs of autosomes and two sex chromosomes (X- and Y-chromosome). The Y-chromosome determines the male sex of its carrier: males carry one X- and one Y-chromosome, while females carry two X-chromosomes. As will be outlined below, the autosomes, sex chromosomes (especially the Y-chromosome), and mtDNA have different modes of inheritance and thus different properties that affect their use in molecular anthropological studies.

The nuclear genome is inherited from both parents: one of each pair of chromosomes comes from the mother, the other from the father. The differences between the autosomes and the sex chromosomes is the provenance of the two chromosomes of each pair: since the Y-chromosome is present only in males, it is always contributed to male offspring by the father, with the X-chromosome in males always coming from the mother. Thus, the Y-chromosome is inherited solely in the paternal line, from fathers to sons; autosomes, in contrast, as well as the X-chromosome in females, are inherited from both parents. The behaviour of mtDNA is complementary to the Y-chromosome: although all humans carry mtDNA, it is passed on only from the mother. We can thus distinguish between the autosomes and the 'uniparental markers' Y-chromosome and mtDNA; the pattern of inheritance of the X-chromosome, which is present in two copies in females and only one in males, is more complicated and will not be considered here. These differences in inheritance between the uniparental markers and the autosomes have important implications for molecular anthropological analyses: the Y-chromosome can help us elucidate the prehistory of the paternal half of a population, the mtDNA can highlight the prehistory of the maternal half, while the autosomes provide insights into the prehistory of the population as a whole. These sex-specific patterns of inheritance of the uniparental markers are very useful when it comes to elucidating the sociocultural aspects of prehistory, and therefore mtDNA and the Y-chromosome remain highly important for studies of population prehistory. However, it should be kept in mind that they provide a relatively restricted view of prehistory: while every man has four grandparents, eight great-grandparents, sixteen great-great-grandparents and so forth, which all contributed to his autosomes and will thus be included in our study of population prehistory when we use autosomes, studying his Y-chromosome includes only one of those sixteen great-great-grandparents (his father's father's father's father), as does studying the mtDNA (the mother's mother's mother's mother). These limitations of the uniparental markers should be kept in mind when evaluating molecular anthropological studies.

In addition to the uniparental vs. biparental mode of inheritance described above, there is yet another important difference between the autosomes and the uniparental markers. This is the process called recombination, which all autosomes undergo during production of germ cells, but which does not take place in mtDNA, and which only affects very small portions of the Y-chromosome. Recombination is the process whereby two matching (homologous) chromosomes align, physically cross over each other, break at the crossing points, and fuse with their counterpart, resulting in molecules that stem from the two separate chromosomes of each pair. This results in enormous variation, ensuring that the chromosomes passed on by an individual will contain random portions of DNA from each of that individual's parents. This means that any newly arising mutation might end up in a very different physical context from that in which it arose. Therefore, it is not possible to reconstruct the chronological sequence in which mutations arose on the autosomes. In contrast, since the uniparental markers do not undergo recombination, they are passed on virtually identical from father to son and from mother to offspring; the only variation stems from newly arising mutations. In

the absence of recombination, these mutations remain in situ, permitting the reconstruction of the chronological order in which they arose: if I find mutations A, B, and C in a subset of individuals in a population, mutations A, B and D in another subset, and mutations A and E in a third subset, then I can reconstruct that mutation A must have happened first, that B and E would have happened next, but independently of each other, and that C and D happened in individuals carrying the B mutation, but not the E mutation. Stated differently, I can reconstruct that individuals carrying mutations C and D share a more recent common ancestor with each other than they do with individuals carrying mutation E. Such groups of related molecules defined by specific diagnostic shared mutations are called *haplogroups* in molecular anthropology – they can be thought of as analogous to subgroups in historical linguistics that are defined by shared innovations. Since the genetic variation in humans is geographically substructured, particular haplogroups can be characteristic of certain geographic regions and sometimes even of particular populations or groups of populations. These are then powerful tools for detecting prehistoric admixture between populations.

The advantages of mtDNA and the Y-chromosome are thus that they allow insights into sex differences in prehistoric demographic processes, such as migration and marriage patterns, and that they allow us to trace mutations back in time and space and thus to straightforwardly detect the source of admixture. The disadvantages of these two markers are that they constitute only a minute portion of the human genome and are thus far more susceptible to random effects (called genetic drift) – for instance, if a man has ten daughters, but no sons, then his Y-chromosomal lineage will die with him (unless he has brothers who have sons) – his autosomes however, will be passed on to the next generation. Furthermore, as explained above, they are informative of only a tiny fraction of the ancestors of each individual. The advantages of autosomal DNA are the wealth of information it carries concerning our genetic prehistory, with an unbiased perspective on all the ancestors of each individual and its correspondingly lower susceptibility to genetic drift effects; with the larger amount of variation carried in our autosomes, estimates of rates and even time of admixture are far more precise. The disadvantage of autosomal DNA is that it does not permit any insights into sex differences in prehistoric events, and cases of prehistoric admixture are not as easily traceable to particular source populations.

Notes

1 I am grateful to the Max Planck Society for providing me with generous funding from 2007–2011 to pursue the interdisciplinary approach to language and population contact presented here with a Max Planck Research Group. Furthermore, I thank Bethwyn Evans and Claire Bowern for comments on a draft of this chapter, as well as Sarah Thomason for comments in response to a presentation at the Workshop 'Foundations of Historical Linguistics' held in Boston, MA, in January 2013. Their input considerably improved this chapter (but of course all errors remain mine).

2 However, currently the language is being given up in favour of Russian – a situation brought about by the exile of the Kalmyk community to Siberia in 1943 (Baranova 2009).

3 In southern Africa one finds two large and quite distinct groups of indigenous languages and peoples: 'Khoisan' and Bantu. The so-called 'Khoisan' languages are spoken by diverse groups of foragers and pastoralists nowadays settled predominantly in Botswana and Namibia. These languages do not form an accepted genealogical unit; instead, specialists agree that they belong to three independent language families (cf. Güldemann 2008). Nevertheless, I will use the cover term 'Khoisan' to refer to these languages and the people speaking them for ease of reference. While the foragers speaking Khoisan languages are widely assumed to be the autochthonous inhabitants

of southern Africa (Deacon and Deacon 1999), the Bantu languages are commonly accepted to be relative newcomers to the area. They are widely associated with the immigration of agriculturalists with Iron Age technology approximately 2000–1200 years before present.

4 Haacke refers mainly to the peripheral (northwestern) dialects of the Damara–Nama language, such as the Sesfontein dialect and especially Hai||om. While the Hai||om (a forager group from north Namibia) are genetically indeed close to other Khoisan, and especially the !Xuun, none of the other Damara subgroups are genetically differentiated in any way; thus, the conclusion that the Damara as a whole shifted to the Khoe language they speak today holds for all of them.

5 While in the Bantu languages of South Africa belonging to the Nguni subgroup a complex taboo system is assumed to have played a major role in the incorporation of click consonants, this cannot have been a causal factor in the southwestern Bantu languages: here, no such system of taboo is known (Bostoen and Sands 2012).

6 Actually, five different mtDNA sequence types are carried by eight Fwe individuals; of these five types, four are so divergent that they must have entered the Fwe gene pool independently of each other (the fifth could have evolved in the Fwe through novel mutations). Four mtDNA types of Khoisan origin might not seem like a lot, and it might therefore be puzzling why from this I conclude that the contact must have been fairly intense. However, it should be noted that we considered only a small subset of individuals from the Fwe population – the sample size was only 33 – and that if our sample was indeed random, as we expect it to be, the number of Khoisan maternal lineages in the population as a whole would correspondingly be larger. Furthermore, with mtDNA analyses we are only able to detect maternal geneflow that resulted in female offspring, so that we are probably underestimating the overall level of Khoisan admixture.

Further reading

Güldemann, Tom and Mark Stoneking. 2008. A historical appraisal of clicks: a linguistic and genetic population perspective. *Annual Review of Anthropology* 37(1): 93–109.

Hunley, Keith and Jeffrey. C. Long. 2005. Gene flow across linguistic boundaries in Native North American populations. *Proceedings of the National Academy of Sciences of the United States of America* 102(5): 1312–1317.

Kayser, Manfred, Ying Choi, Mannis van Oven, Stefano Mona, Silke Brauer, Ronald J. Trent, Dagwin Suarkia, Wulf Schiefenhövel and Mark Stoneking. 2008. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. *Molecular Biology and Evolution* 25(7): 1362–1374.

Pakendorf, Brigitte, Cesare de Filippo and Koen Bostoen. 2011. Molecular perspectives on the Bantu expansion: a synthesis. *Language Dynamics and Change*, 1(1): 50–88.

References

Aikhenvald, Alexandra Y. 2002. *Language contact in Amazonia.* Oxford: Oxford University Press.

Baranova, V. V. 2009. Jazykovaja situacija v Kalmykii: sociolingvističeskij očerk. In S. S. Saj, V. Baranova and N. V. Serdobol'skaja (eds) *Issledovanija po grammatike kalmyckogo jazyka*. Acta Linguistica Petropolitana. Trudy instituta lingvističeskix issledovanij. St Petersburg: Nauka, 22–41.

Barbieri, Chiara, Mark Whitten, Klaus Beyer, Henning Schreiber, Mingkun Li and Brigitte Pakendorf. 2012. Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. *Molecular Biology and Evolution* 29(4): 1213–1223.

Barbieri, Chiara, Anne Butthof, Koen Bostoen and Brigitte Pakendorf. 2013a. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. *European Journal of Human Genetics* 21: 430–436.

Barbieri, Chiara, Mario Vicente, Jorge Rocha, Sununguko W. Mpoloka, Mark Stoneking and Brigitte Pakendorf. 2013b. Ancient substructure in early mtDNA lineages of southern Africa. *American Journal of Human Genetics* 92(2): 285–292.

Barbieri, Chiara, Tom Güldemann, Christfried Naumann, Linda Gerlach, Falko Berthold, Hirosi Nakagawa, Sununguko W. Mpoloka, Mark Stoneking and Brigitte Pakendorf, 2014. Unraveling the complex maternal history of southern African Khoisan populations. *American Journal of Physical Anthropology* 153: 435–448.

Barbujani, Guido. 1991. What do languages tell us about human microevolution? *Trends in Ecology and Evolution* 6(5): 151–156.

Belle, Elise M. S. and Guido Barbujani. 2007. Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. *American Journal of Physical Anthropology* 133(4): 1137–1146.

Berniell-Lee, Gemma, Francesc Calafell, Elena Bosch, Evelyne Heyer, Lucas Sica, Patrick Mougiama-Daouda, Lolke van der Veen, Jean-Marie Hombert, Lluis Quintana-Murci and David Comas. 2009. Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Molecular Biology and Evolution* 26(7): 1581–1589.

Beyer, Klaus and Henning Schreiber. 2013. Intermingling speech groups: morpho-syntactic outcomes of language contact in a linguistic area in Burkina Faso, West Africa. In Isabelle Léglise and Claudine Chamoreau (eds) *The interplay of variation and change in contact settings*. Amsterdam: John Benjamins, 107–134.

Bläsing, Uwe 2003. Kalmuck. In Juha Janhunen (ed.) *The Mongolic languages*. London: Routledge, 229–247.

Bostoen, Koen and Bonny Sands. 2012. Clicks in south-western Bantu languages: contact-induced vs. language-internal lexical change. In Matthias Brenzinger (ed.) *Proceedings of the 6th World Congress of African Linguistics Cologne 2009*. Köln: Rüdiger Köppe Verlag, 129–140.

Deacon, H. J. and Janette Deacon. 1999. *Human beginnings in South Africa: uncovering the secrets of the Stone Age*. Walnut Creek, California: Rowman Altamira.

de Filippo, Cesare, Koen Bostoen, Mark Stoneking and Brigitte Pakendorf. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings. Biological sciences / The Royal Society* 279(1741): 3256–3263.

Forster, Peter and Colin Renfrew. 2011. Evolution. Mother tongue and Y chromosomes. *Science* 333(6048): 1390–1391.

Güldemann, Tom. 2008. A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. *Southern African Humanities* 20: 93–132.

Gumperz, John J. and Robert Wilson. 1971. Convergence and creolization. A case from the Indo-Aryan/Dravidian border in India. In Dell Hymes (ed.) *Pidginization and Creolization of Languages. Proceedings of a Conference held at the University of the West Indies Mona, Jamaica, April 1968.* Cambridge: Cambridge University Press, 151–167.

Haacke, Wilfrid H. G. 2008. Linguistic hypotheses on the origin of Namibian Khoekhoe speakers. *Southern African Humanities* 20: 163–177.

Hunley, Keith, Michael Dunn, Eva Lindström, Ger Reesink, Angela Terrill, Meghan Healy, George Koki, Françoise R. Friedlaender and Jonathan S. Friedlaender. 2008. Genetic and linguistic coevolution in northern Island Melanesia. *PLoS Genetics* 4(10): e1000239.

Hunley, Keith, Claire Bowern and Meghan Healy. 2012. Rejection of a serial founder effects model of genetic and linguistic coevolution. *Proceedings. Biological Sciences / The Royal Society*, 279(1736): 2281–2288.

Lansing, J.Stephen, Murray P. Cox, Sean S. Downey, Brandon M. Gabler, Brian Hallmark, Tatiana M. Karafet, Peter Norquest, John Schoenfelder, Herawati Sudoyo, Joseph C. Watkins and Michael F. Hammer. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences of the United States of America* 104(41): 16022–16026.

McMahon, Robert. 2004. Genes and languages. *Community Genetics* 7(1): 2–13.

Nasidze, Ivan, Dominique Quinque, Manijeh Rahmani, Seyed Ali Alemohamad and Mark Stoneking. 2006. Concomitant replacement of language and mtDNA in South Caspian populations of Iran. *Current Biology* 16(7): 668–673.

Nasidze, Ivan, Dominique Quinque, Isabelle Dupanloup, Richard Cordaux, Lyudmila Kokshunova and Mark Stoneking. 2005. Genetic evidence for the Mongolian ancestry of Kalmyks. *American Journal of Physical Anthropology*, 128(4): 846–854.

Pakendorf, Brigitte. 2007. *Contact in the prehistory of the Sakha (Yakuts): linguistic and genetic perspectives*, Utrecht: LOT.

Pakendorf, Brigitte, Innokentij N. Novgorodov, Vladimir L. Osakovskiy, Al'bina P. Danilova, Artur P. Protod'jakonov and Mark Stoneking. 2006. Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. *Human Genetics*, 120(3): 334–353.

Pakendorf, Brigitte and Innokentij N. Novgorodov. 2009. Loanwords in Sakha (Yakut), a Turkic language of Siberia. In Martin Haspelmath and Uri Tadmor (eds) *Loanwords in the world's languages: a comparative handbook*. Berlin: de Gruyter Mouton, 496–524.

Pickrell, Joseph K., Nick Patterson, Chiara Barbieri, Falko Berthold, Linda Gerlach, Mark Lipson, Po-Ru Loh, Tom Güldemann, Blesswell Kure, Sununguko Wata Mpoloka, Hirosi Nakagawa, Christfried Naumann, Joanna Mountain, Carlos Bustamante, Bonnie Berger, Brenna Henn, Mark Stoneking, David Reich and Brigitte Pakendorf. 2012. The genetic prehistory of southern Africa. *Nature Communications* 3: 1143; doi:10.1038/ncomms2140.

Poloni, E. S., O. Semino, G. Passarino, A. S. Santachiar-Benerecetti, I. Dupanloup, A. Langaney and L. Excoffier. 1997. Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *American Journal of Human Genetics* 61(5): 1015–1035.

Quintana-Murci, Lluís, Hélène Quach, Christine Harmant, Francesca Luca, Blandine Massonnet, Etienne Patin, Lucas Sica, Patrick Mouguiama-Daouda, David Comas, Shay Tsur, Oleg Balanovsky, Kenneth K. Kidd, Judith R. Kidd, Lolke van der Veen, Jean-Marie Hombert, Antoine Gessain, Paul Verdu, Alain Froment, Serge Bahuchet, Evelyne Heyer, Jean Dausset, Antonio Salas and Doron M. Behar. 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5): 1596–1601.

Ross, Malcolm. 1996. Contact-induced change and the comparative method: cases from Papua New Guinea. In Mark Durie and Malcolm Ross (eds), *The Comparative Method reviewed. Regularity and irregularity in language change*. New York, Oxford: Oxford University Press, 180–217

——2003. Diagnosing prehistoric language contact. In Raymond Hickey (ed.) *Motives for language change*. Cambridge: Cambridge University Press, 174–198.

——2007. Calquing and Metatypy. *Journal of Language Contact* 1(1): 116–143.

——2013. Diagnosing contact processes from their outcomes: the importance of life stages. *Journal of Language Contact* 6(1): 5–47.

Ruhlen, Merritt. 1987. *A guide to the world's languages.* Volume 1: *Classification*. Stanford: Stanford University Press.

Stoneking, Mark. to appear. *Introduction to Molecular Anthropology*. Malden: Wiley Blackwell.

Thomason, Sarah G. and Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Whitten, Mark, Mingkun Li, Connie Mulligan, Njamhishig Sambuughin, Evelyne Heyer, Laure Segurel, Victor Wiebe, Michael Crawford, M.J. Mosher, Sergey Makarov, Victor Spitsyn, Innokentij Novgorodov, Vladimir Osakovsky and Brigitte Pakendorf. in preparation. Uncovering demographic histories of Siberian populations using complete mtDNA genomes.

Wood, Elizabeth T., Daryn A. Stover, Christopher Ehret, Giovanni Destro-Bisol, Gabriella Spedini, Howard McLeod, Leslie Louie, Mike Bamshad, Beverly I. Strassmann, Himla Soodyall and Michael F. Hammer. 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *European Journal of Human Genetics*, 13(7): 867–876.