Roma: linguistic archaeology of nomads

PETER BAKKER & ASTRID MONRAD

Bakker, P. & Monrad, A. 2011. *Roma: linguistic archaeology of nomads.* AmS-Varia 53, 35–44. Stavanger. ISSN 0332-6306, ISBN 978-82-7760-152-6, UDK 314.

This paper promotes the possibility of using linguistic data for reconstructing historical events, in particular migrations. It deals specifically with the migrations of the ancestors of the Roma/Gypsies from India to Europe at least around a millennium years ago. Their migration route has been reconstructed on the basis of data on sound changes and borrowings in the language of the Gypsies, Romani, despite the complete lack of archaeological data and the scarcity of historical documentation. The introduction of words and structures from Dardic languages of North India, Iranian languages, Armenian, Greek and South Slavic languages into Romani, can be mapped into a fairly specific migration route. The study of changes in Romani that are shared with other documented languages from Central and South Asia that can be dated, makes it possible to tentatively attach dates to the migrations, and possibly impression of length of stay. Also the lack of loanwords from Arabic, which otherwise penetrated Western Asian languages from ca. 750 A.D., sheds light on the chronology. In recent years, molecular genetic research on the genetic profile of the Roma has confirmed the linguistic data. The authors argue that linguistic data can be handled as symbolic archaeological artefacts, and thus shed light on historical events.

Keywords: Roma, Gypsies, linguistic archaeology, language reconstruction, historical linguistics, genetics, dating, migration, lexicon, borrowing, nomads

Peter Bakker, Linguistics, Department of Communication and Aesthetics, Aarhus University.
Nordre Ringgade 1, 8000 AARHUS C, DENMARK. Telephone: (+45) 89426553. Telefax: (+45) 89426570. E-mail: linpb@hum.au.dk
Astrid Monrad, Research Center for Grammar and Language Use, Aarhus University
Nordre Ringgade 1, 8000 AARHUS C, DENMARK. E-mail: astrid.monr@gmail.com

1. Introduction

The history of the Roma (Gypsies) constitutes a prime example of a case in which linguistic data can be used to establish knowledge about the history and prehistory of a population group. Especially in the past decade there is an increase of literature in which linguistic data are used for reconstruction and dating of population movements, for instance recent work on Austronesian migrations (Gray & Jordan 2000, Greenhill *et al*. 2010). In many cases linguistic data and archaeological data complement one another, or can be used as an argument in archaeological discussions (e.g. Gray & Atkinson 2003). On the other hand, everything that is known about the prehistory of the Roma, and much of their history as well, is based on the study of their language, Romani. Only recently molecular genetic data have been collected as well. This paper will show that these linguistic data are a reliable (within certain margins, of course) indicator of historical events. In fact, linguistic data may be more reliable than historical, written sources, since written sources can be subject to lying, and authors may have certain personal, social or political interests to distort the truth. Both linguistic facts and archeological data are subject to interpretations by scholars, and interpretations of linguistic data are not more speculative than those made by archaeologists.

In this paper we will first introduce the Roma (section 2), then we will give a brief and simplified introduction to the relevant methods used in historical linguistics - that branch of linguistics that deals with the reconstruction of earlier stages of languages, and with principles of language change (section 3). After that we will discuss the history of the Roma (section 4) and focus on the migration route such as can be deducted from linguistic data (section 5). We will present our conclusions in section 6.

2. Who are the Roma?

The Roma are a heterogenous ethnic group, known by outsiders under names such as Gypsies, Tsiganes, Zigeuner, and more recently also as Roma and Sinti (Tcherenkov & Laederich 2004, Liégeois 2008). The last two are used in their own language as their labels for themselves. There are around six million people

mostly in (Eastern) Europe. In the popular image, the Roma are considered nomads, and this is also part of the self-image of many Roma. In reality, however, the vast majority of Roma have been settled for more than 200 years. This rather misleading image may have been caused by the fact that the Roma indeed came from somewhere else some time in the past, and the fact that some groups or families indeed travelled in the recent past, either permanently, or temporarily, often for purposes of business.

Culturally, we can point at strong in-group attitudes and strong (extended) family bonds, leading to a continuum of solidarity from the family to the clan or the professional group, to wider circles (Liégeois 2011).

Outsiders often use the term Gypsy or Gypsies to refer to a range of individuals and groups, sometimes based on their lifestyle, often based on a supposed nomadic background. Not all of the people thus classified would fall under the label "Gypsies" or "Roma" as used throughout this paper, and as used by Roma themselves. This paper deals with those people who identify as Gypsies, and who speak (or whose ancestors spoke) a language they call "Romanes" (cf. Bakker *et al.* 2000). We are dealing with the following extremes of definitions of Gypsies: on the one hand, any nomadic group or person (cf. "Gypsy scholar", for an academic person who moves from place to place), on the other hand those who speak a language called Romanes, and who are born into specific families. The words Rom, Romani and Romanes should not be confused with Romanian and Romania – the similarity in form is coincidental.

The label **Gypsies** is derived from the word "Egyptian", just like the Basque label "Ijito", Greek "Jifti", Spanish "Gitano" and French "Gitane", as early Gypsies reported they came from "Little Egypt". This place has never been identified or associated with a specific location in or outside Egypt. An Egyptian origin of the Roma can be ruled out, among others on linguistic grounds.

Other frequent outside labels include a set of terms supposedly derived from a Greek word **tsinganos**, perhaps "untouchable". These terms include Germanic "Zigeuner", French "Tsigane", Turkish "Çingene". There are also a few labels relating to geographical areas, such as "Tatere" (Scandinavia, North Germany) and "Bohémiens" (France).

Insiders labels are: **Rom** "human, man"(also, derived from that, Romano, Romanichel), from Indic *Dom* and /ḍom/"man, caste/ethnic group of smiths and musicians", the retroflex /ḍ/ developed into a lateral in Europe; Matras 2002:36, 40; see also section 5.1); **Sinti**, of unknown origin (in any case not cognate with *Sindh* or *Hindi*), **Manouche**, from Romani *manuš* "person", **Calo/Kaalo** from Romani *kalo* "black" in western Europe (Spain, Wales, Scandinavia). The Sinti live in Germany and adjacent countries, as well as Northern Italy, Slovenia, Hungary and the Soviet Union (Tchernkov & Laederich 2004, Liegeois 2008).

All these groups, even if they do not call themselves Roma, call their language Romanes, and the language is indeed the same, albeit split into a range of dialects.

The Roma also have a number of subgroups, almost always named after a profession, or after a country or region. Best known to outsiders among the occupational names are **Lovari**, from Hungarian *lo* "horse", and **Kalderash** from Roumanian *câldâr* "kettle". Group names based on country or region are for example **Olah/Vlah** "Roumania", **Serbika** "Serbia" and **Rom-Ungro** "Hungary".

The genetic affiliation of the Romani language is Indic. This means that the basic vocabulary and grammatical structures are derived from Indic languages, and ultimately go back to Sanskrit. Some examples: Romani *andro, arno*, Sanskrit *ā<u>nd</u>a* "egg", R. *mar-* "hit, kill", S. *mārayati* "to kill", R. *mas*, S. *mā<u>m</u>sa* "meat". Sanskrit and the Indic group of languages belong to the Indo-European language family, to which also Norwegian and English belong. Norwegian and English are part of the Germanic branch. The Romani numerals *duj* and *trin* (Sanskrit dvi tri), for instance, are clearly cognate with the Norwegian and English numerals *to/tre, two/three*. Ultimately Norwegian and Indic languages like Romani descend from a putative language called Proto-Indo-European, spoken somewhere in Asia between 5,000 and 10,000 years ago.

The Indic connection of Romani was discovered by several people independently in the 1780s in Britain (Bryant; Sampson 1911), in Hungary/the Netherlands (Vali/Valyi; Hancock 1993; see also Willems 1997), Russia (Bacmeister, Catherine the Great; Matras 2003:2) and Germany (Rüdiger; Matras 1999). The genealogical affiliation was most convincingly shown by J. Rüdiger in Germany, who pointed to several lexical and grammatical similarities between Hindi and Sinti Romani (Matras 1999). Whereas others like Bryant and Bacmeister had drawn attention to similarities in the lexicon, just words are insufficient as proof. Rüdiger was the first to point out the grammatical similarities: "As regards the grammatical part of the language the correspondence is no less conspicuous, which is an even more important proof of the close relation between the languages.» (translation from Rüdiger's German in Matras 1999:95). Rüdiger points to similarities in grammatical forms such deictics, adverbs, prepositions and personal pronouns, including case endings,

such as Romani *tuke*, Hindustani *tumku* "for you". Among the lexical similarities discussed by Rüdiger we can mention Romani *anter tschutschi* "in breasts", Hindustani *schischi anter* (same meaning; note the shifted word order, rightly attributed to contact with European languages by Rüdiger).

3. Historical linguistics

It would be fair to state that Rüdiger discovered the Indic genetic affiliation of Romani. The question to discuss here is how do linguists decide that two languages are "genetically related"? I.e. how do they decide that two languages have a common origin?

In order to prove this, three basic conditions have to be met. First, there have to be similarities in common **vocabulary** (body parts, family members, weather, etc.), both in their *form and meaning*. If both of these are similar, it can be assumed that the words reflect the same earlier word, and can be said to be cognates. It is not difficult to see that Danish *hjerte*, Dutch *hart* and German *Herz* are cognate with the English word *heart* of the same meaning, but it needs more sophistication and more data are necessary from other languages to prove that French *coeur* and Romani *ilo* are also cognates - which they indeed are. The reconstructed Proto-Indo-European root of "heart" is ****k'r.d-.*** If one has a range of cognate words from the same set of languages, one can try to reconstruct the original sounds of the word from which all these words derive, and this technique has proven successful in many ways.

In the case of the word for "heart", the meaning has not changed, only the form, but in other cases we may notice a change of meaning as well. Here, things may get more complicated, as in the cognates Danish *tømrer* "carpenter", English *timber* "wood", German *Zimmer* "room", Dutch *timmeren* "to hammer", where the meanings are quite different, but the words share a clear core: something to do with wood and the work with it.

Just one set of words similar in form and meaning, however, is not enough. It can be chance, as English *dog* and *dog* with the same meaning in Mbabaram, an Australian Aboriginal language, or Hawaiian *aeto* "eagle" and classical Greek *aetos* "eagle". In other cases formal similarities can be independent results of sound imitation, like the Japanese write the sound roosters make as *kokekokkoo*, and the Dutch as *kukeleku*.

Similarities can also be based on universals of ease of articulation: many languages have words for "father" and "mother" based on the most contrastive consonants and vowels. By a subsequent combination of the maximum closure of the mouth, yielding a /p/ or an /m/ sound, with a maximum opening of the mouth producing an "Aah" sound (/a/), one gets words like "mama" and "papa", almost universal for either "father" or "mother".

Finally, words can be taken over from one language into another. Finnish and English share words like *gorilla* and *president(ti)*, but these are borrowed words in both languages.

Even though most of these examples of lexical similarities that are not based on a shared genetic origin seem trivial, it is not always easy to distinguish inherited words from borrowings, sound imitations, or chance. Furthermore, shared words are not enough, since languages can borrow close to half of their vocabulary, even those words used on a daily basis (English words like *kid, husband, take* and *they* are all borrowings from Scandinavian). It will be clear that a significant number of words is common, and not only that, these words must also display regular sound changes. For example Spanish *cabro* and French *chèvre* can be proven to be cognates because words pairs like *castello/chateau* and *cabeza/chef* show similar sound changes. Where French has the sound /š/ as in English *ship*, Spanish has /k/ (written <c>), and where French has /ɛ/, Spanish has /a/, and where French has /f/ or /v/, Spanish has /b/. Sound changes in these cognate words must be recurring and **regular.**

In order to prove a genetic relationship, however, also similarities in **grammatical** elements (pronouns, verbal and nominal endings, etc.) are necessary, both in *form and meaning.* Some categories of free grammatical elements are relatively stable, in the sense that they are rarely borrowed, but on the other hand they appear more unstable, because they may be reduced in form through frequent use, and subsequently replaced by more emphatic forms. Thus, the Scandinavian articles *den, det, -en, -et* were originally demonstratives, and new demonstratives (e.g. *denne, dette*) were formed afterwards. Morphological endings, especially inflectional elements indicating grammatical relations, tend not to change fast. All these factors are relative: anything can change. But if one can demonstrate similarities in form and meaning in the lexicon and in the morphology, and regular sound changes in these two languages, these languages are genetically related.

3.1. Dating in historical linguistics

Before going on with Romani, we have to mention dating techniques used by historical linguists. These are of two types: relative dating and absolute dating. The first means that one can date changes relative to one another (e.g. first X changed to Y, and then Y changed to A, or Z changed to A), and the second means that one can attach a date to a certain change. For example, if we

have the forms *jimansh* /djimãš/ in a French variety for standard French *dimanche* /dimãš/ "Sunday", showing a change of /d/ in front of the vowel <i> whereas this did not happen in words like *difand /difâd* "defend" (standard French *défend*), we can conclude that /e/ <é> changed to /i/ in words like *défend* after /d/ changed to /dj/, otherwise we would have the form /djifâd/.

Absolute dating can be established if we have a series of dated documents or inscriptions, and we can observe a certain element before and after its change - the dates will enable us to date the changes in the language. These changes do not take place overnight, but sometimes over several centuries. From the time a changed form is first used by one individual speaker, it has to spread through the language to other, similar words, and it has to spread through the community to other speakers.

Linguistic changes include borrowings of words and structures. Since words are usually borrowed from languages of more powerful groups into other languages, one can draw conclusions about social and power relations between groups in former times. Borrowing can sometimes also be used in dating.

There are also more controversial techniques, in which one assumes a certain constant rate of change. The technique of glottochronology is one of those: assuming that languages replace a certain proportion of their vocabularies with a fixed regularity, one would be able to date changes. The latter technique is rather controversial among linguists, but it is being used both by linguists and nonlinguists (see some of the contributions in Renfrew *et al*. 2000 for a range of viewpoints). Recently, more precise dating techniques have been proposed, but it is still too early to know whether these will prove acceptable (Greenhill *et al*. 2010).

More information on these techniques can be found in e.g. Campbell (1998).

3.2. Romani as an Indic language

The statement that Romani is an Indic language, is based on those criteria established in historical linguistics for the establishment of a genetic connection. The core vocabulary of Romani is Indic: many hundreds of words are closely cognate with Modern, Middle and/or Old Indic words (Matras 2002:20–30). Most Romani kinship terms, body part terms, weather words, basic verbs, lower numerals, etc. can be related to Indic languages, including some verbs with irregular past tenses. These irregularities are in themselves proof that the Romani vocabulary is inherited from Indic, since such irregularities are never borrowed.

Furthermore, there are many regular sound correspondences between these Indic elements in Romani and languages of India (see Matras 2002:30–41). At the grammatical level, almost all grammatical endings are Indic, or in fact those connected to the pre-European vocabulary: the case endings, plural endings, verbal inflection and derivation (Matras 2002:42–45). For any linguist, these facts can only lead to the conclusion that Romani is an Indic language.

Romani does not only consist of Indic elements, however, and it is especially those elements that allow us to deduct information about the early history of the ancestors of the Roma.

4. Where do the Roma come from?

If Romani is an Indic language, do the Roma, or rather their ancestors, come from India? Of course there is no direct proof for such a claim. There is, however, a huge amount of circumstantial evidence. For instance physical features of the Roma, who are virtually always darker than the surrounding populations. This impression has been corroborated by studies of genetic data (see below). Cultural data may also shed light on this, but they should be used with caution.

4.1. Genetic data

Lately, geneticists have started showing an interest in the Roma, and findings within this field support linguistic theories about the history of the Roma. Several studies on human DNA (see for example Gresham *et al*. (2001), Ioviță & Schurr (2004), Morar *et al*. (2004), Kalaydjieva *et al*. (2005)) show a genetic link between India and the Roma. The Roma have genetic features in common with South Asian populations that are not found in other Europeans.

A number of studies on hereditary diseases have shown so-called founder effects of a population, indicating that the Roma descend from a small group of ancestors (e.g. Navarro & Teijeira 2003, Kalaydjieva *et al*. 2005, Bouwer *et al*. 2007). Some subgroups as well point to common descent, for instance because certain hereditary diseases are limited to Roma, or to Roma belonging to the Vlax subgroup.

These studies have, directly or indirectly, been inspired by linguistic data, and confirmed that not only the language of the Roma, but also the population itself, has an origin in India. The fact that founder effects have been found, supports the theory that the Roma left India as a small and distinct group, and that the Roma are neither from Egypt, as legend has it, or that Roma would be Europeans who emerged as a group through social stigmatisation, as has also been suggested.

Morar *et al*. (2004), for instance, found that "the identity of the congenital myasthenia 1267delG mutation

in Gypsy and Indian/Pakistani chromosomes provided the best evidence yet of the Indian origins of the Gypsies", as it is limited to these isolated population groups. They conclude that the "entire Gypsy population was founded approximately 32–40 generations" ago.

However, we want to focus on non-biological evidence for the origin of the Roma here. For a more detailed summary of genetic studies of the Roma, linked with linguistics, see Bakker (in press).

4.2. Historical, cultural and archaeological evidence for the Indic origin of the Roma

In contrast to linguistic and biological evidence indicating an Indic origin of the Roma, and the migration between South Asia and the arrival in Southeastern Europe in or around the 13th century, there is surprisingly little.

As far as we know, there is not a single archaeological study of the migration of the ancestors of the Roma from India to Europe – but this absence of evidence of course does not mean that such a migration has not taken place. It seems inherently more difficult to study the archaeology of nomads than of other groups, especially groups such as the Roma (assuming that their ancestors had life styles and occupations similar to many in the recent history of Europe), who always live separated from other groups, but are nevertheless dependent on them. In any case, archaeology seems to be of no help for a study of Gypsy origins.

Written history does not provide a clue either. If the ancestors of the Roma travelled overland from India to Europe sometimes before the 12th century, they came from an area with an old written tradition (a.o. Sanskrit, Apabhramsa, Prakrit, etc.), and crossed through territories with literate populations: Arabic script was used from the 6th century, Armenian from the 5th century, Georgian from ca. the 12th century) and Byzantine Greek from the 6th century. However, very few historical documents have been unearthed relating to populations who could have been the ancestors of the Roma, and none of these can be connected to them with any degree of certainty (see Matras 2002).

Sometimes cultural similarities between Roma and Indian populations are mentioned, such as the importance of ritual purity, food taboos and hygiene, plus an organization along caste-like professional lines among Roma as among South Asian populations, but these can just as well be attributed to chance, since they often differ in their specifics, and there are important differences between Roma groups as well.

What evidence is there for an Indian origin of the Roma? The hardest pieces of evidence are the linguistic and molecular genetic data discussed above. There are absolutely no convincing and unambiguous historical or archaeological data for this claim, however. In the remainder of this paper, we will focus on the reconstruction of the migration route of the ancestors of the Roma from India to southeastern Europe, based exclusively on linguistic evidence.

5. Migration route

We know from written historical sources that Roma were first unambigously mentioned in southeastern Europe in the early 14th century. There are some more dubious older sources as well (see Gilsenbach (1998) for a overview, and Tcherenkov & Laederich (2005) for more critical use of these early sources). It is not known exactly when and why the ancestors of the Roma left India. In this section we will summarize the main findings on the migration route between India and Europe.

In trying to reconstruct the migration route between India and Europe, there is no historical evidence at all. No chronicler mentioned the observation of, say, 30,000 people travelling westwards through Asia. Neither did these people leave any identifiable traces that archaeologists have laid their hands on, and identified as being "Gypsy". We have to rely on some limited genetic data (which we will largely ignore here), but on the other hand we have ample linguistic data, that allow us to reconstruct this migration route and date this to a rather detailed level. Linguistic facts of the Romani language can be considered to be equivalent to the non-material artefacts enabling us to practice linguistic archaeology.

For students of archaeology or history, it may seem odd to adduce linguistic data to reconstruct historical events, but in some sense linguistic evidence may be stronger than historical sources: authors of documents may intentionally distort facts, but no one intentionally changes sounds or structures in a language in order to mislead interpreters of those: linguistic data therefore do **not lie** (in contrast to written documents).

In addition to the methodological principles outlined above in section (3), we should add here that shared historical linguistic data as such do not prove a lot. Shared elements preserved from an earlier period (conservatisms) in themselves do not reveal much in the way of historical data: shared innovations reveal much more about historical events. The fact that the word *bal(a)* would mean "hair" in Sanskrit, Hindi and Romani, does not provide any indication regarding historical developments between the stage of Old Indic and the present day. However, when two languages, or two varieties of a language, share a particular innovative development not found in other varieties, this

would indicate that these innovators share some of their later history with each other, and that they were separated from other groups who did not innovate. When innovations can be shown to have taken place under the influence of other languages, because they share the same unusual developments, one can even point to a certain geographical area or a certain period when these changes took place. In other words, one has to look for shared changes/innovations in order to make meaningful statements about developments in the past, more specifically a migration.

5.1. Shared sound changes

An example of a shared innovation could be the development of the South Indian retroflex /r/, found a.o. in Sanskrit. It is an /r/ produced with the top of the tongue backwards, as in English and Indian languages. Where Sanskrit has this sound in a range of words, Romani has just /r/, /nr/ or even /ndr/, depending on the dialect. Only a few isolated dialects in Turkey and Bulgaria preserved a retroflex sound. These can be plotted on a map of European Romani dialects, showing clear geographical patterns (see Matras 2002:216). When Romani lost the retroflexes, is impossible to date.

5.2. Shared lexical borrowings

Languages can borrow words from other languages. Usually languages borrow from socially and economically dominant languages. In the case of Romani, one can deduct from the presence of a set of loanwords that can be linked to an identifiable language in a Romani variety, that ancestors of these speakers used to live an area where that other language was spoken.

Loanwords found in all branches of Romani include: *baxt* "luck" *ambrol* "pear" from Persian (*baxt*, *amrūd*); *sir* "garlic" and *vazd-* "raise" from Kurdish; *grast* "horse", *kotor* "piece" from Armenian (*grast, kotor*); *khilav* "plum" (Georgian *khliavi*); *drom* "road" *foro* "city" from Greek (*dromos*, *foros* "market"). Loanwords from European languages, including Hungarian, Slavonic and Rumanian, are only found in a limited number of varieties, and are therefore not the result of a shared history. The presence of Rumanian loanwords, for instance, would point to an earlier place of settlement (or colony) of Romani speakers among speakers of Rumanian.

5.3. Diffusion: shared grammatical features from neighboring languages

The same technique can be used for grammatical innovations. Just like languages can take over words and forms, they can take over constructions from other languages, for instance word order, or a grammatical category, e.g. the marking of definiteness. This is just as common as sound changes and lexical borrowing, but less well studied (Heine & Kuteva 2005). Romani is significantly influenced by Balkan languages, notably Greek,

5.4. Shared innovations: caution

All this does not mean that one can just point out similar developments in two languages and then conclude that the languages share some of their history. There are a number of pitfalls, and we discuss the most important ones here. First, shared sound changes can be independent developments. Some sound changes are so natural that they recur in many areas of the world. Experienced linguists, however, know what kinds of sound changes are common and which ones are rare. Also, similar grammatical changes can take place independently in different areas.

Second, shared lexical borrowings may not be the consequence of direct contact between two language groups, but they can come from an intermediary language. The presence of Amerindian words like "tipi" and "wigwam" in Danish does not prove contact between Danes and North American natives: they came into Danish via English.

In cases like this, a certain quantity of shared elements, or a cluster of recurring structures, or cross-linguistically rare phenomena, or a combination of those, are needed in order to provide convincing proof of historical contact.

5.5. Common lexicon of Romani

Having said this, we can return to the lexicon of Romani. The basic lexicon of all Romani varieties is the same for all dialects, from Iran to Wales, and from southern Italy to Finland.

The oldest layer is Indic, and most similar to the languages currently spoken in Central India. There are many hundreds of these words. In addition to that, there are a few words in Romani that are more typical for the so-called Dardic languages, spoken in the mountainous area in and around northern Pakistan. Further there are a significant number of words from Iranian languages such as Farsi/Persian and Kurdish (between 40 and 100). In addition, there are a number of words from languages spoken in the southern Caucasus and North Eastern Turkey, notably Armenian (several dozens, maybe up to 50) and Georgian (a few). The next layer is Greek, with up to 200 words across dialects. Greek is important because of the high number of words (Grant 2003), and the grammatical impact it had as well. The Indic words and the words borrowed from these languages, belong to a fixed set

of somewhat more than 1000 words. Finally, there are some South Slavic words in all Romani varieties, but it can be questioned whether these words are independently borrowed, or that there is a shared set of words that once was common to all varieties. It happens only very rarely that new words from these languages are "discovered" in Romani varieties. An excellent overview of etymologies in one Romani variety can be found in Elśik (2009 n.d.).

The presence of Indic, Dardic, Iranian, Armenian, Georgian and Greek words, suggests a migration route through Asia, through the regions where these languages are spoken.

5.6. Absence of borrowings from Arabic

It should also be remarked that the absence of words from certain languages is significant as well. Whereas virtually all languages spoken in and between Turkey and Pakistan abound with direct and indirect Arabic loans (e.g. via Persian into Turkish), some of them dating from the time of the spread of Islam in the area from ca. 750 A.D., Romani has no Arabic loans, suggesting that Romani speakers somehow missed the spread of Islam and associated terminology through Asia (Matras 2002: 25), either because of extreme social isolation, or because their migration preceded the spread of Islam and Arabic.

5.7. Do these borrowings reflect a migration route?

All this suggests strongly that the ancestors of the Roma left Central India, spent some time in Northern India where they were in contact with speakers of Dardic languages, and travelled through Iranian speaking areas and Armenia to an area south of the Caucasus, and then Greek speaking territories (perhaps in what is now Turkey). They must have travelled in one group, otherwise one cannot explain that all Roma groups all over Europe and beyond, would have borrowed the same, limited set of words from these languages, some as common as "road" (*drom* < Greek), "luck" (*baxt* < Persian) "soul" (*odji* < Armenian) and some very specific ones "eyebrow" (*cimcali* < Georgian), "father of the bride/groom" (*xanamik* < Armenian). Note that these borrowings cannot be explained by the presence of a third source language that happens to contain all these languages.

If we plot the sources of these loanwords on a map of Eurasia, we can notice that they are found more or less on a straight line from India to South Eastern Europe. The most obvious interpretation of these data is that this was the route of immigration from Central India (from where the clearly oldest lexical layer originates) through Asia to southeast Europe. This is most likely correct, but it may be worth pointing out that words can travel easily between languages. In order to prove the migration route, we need additional evidence. We will present some of that below, in the section on grammatical borrowing, which is much less conscious, and which appears to coincide with the lexical borrowing.

Even if we are able to reconstruct a migration route, that does not mean that all problems have been solved. There are several remaining questions, for which a range of answers have been proposed. These questions are: When did the ancestors of the Roma leave India? Why did they leave India? DID the ancestors of the Roma come from India? Why are there no Arabic loans in Romani? Where were the Greek words borrowed? Can we date the presence of the ancestors of the Roma in certain language communities in Asia? These questions will be partly answered.

In this section, the linguistic data are given only in a cursory way. For more details see especially Matras (2002) and Tcherenkov & Laederich (2005).

5.8. Written language sources I: India

The regions from which the ancestors of the Roma originate, and the countries they travelled through, were inhabited by people who had written forms of their languages, and in fact long traditions of literacy. Thanks to those it is possible to track changes in the local languages, as we know when certain changes took place, and in some cases we know that Romani underwent the same change.

India is a case in point, with a documented language history of some 3500 years. The history of the Indic or Aryan languages of India is conveniently divided into three periods:

– *Old* Indo-Aryan: 1500 BC–600 BC
– *Middle* Indo-Aryan: 600 BC–AD 1000
– *New* Indo-Aryan: AD 1000–now

Historically, Romani is most closely related to the so-called Central languages of the Indic branch, such as Hindi, Panjabi, Gujarati, Rajasthani (cf. Masica 1991:appendix II). It has been established that not only the most basic lexicon, but also grammatical traits of Romani are most closely related to Central Indic languages, and these can be traced back ultimately to OIA. Most of the morphological endings are from Central Indic languages (case endings, verbal inflection, verbal derivation, the layered structure of the noun phrase), there are Indic phonemes (most dialects have inherited Indic aspirate stops, a few also retroflex stops). This points to Central India as the cradle of Romani.

We can look at some of the changes that have taken place in the Indic component of Romani and compare those with developments in Indic languages that we

can date because we have written and dated or datable sources for Indic languages for 2500 years. If Romani shares innovations that took place at or before a certain date in India, we can reasonably assume that the ancestors of the Roma were still in South Asia.

We have no space here to go into detail, and refer to Matras (2002:chapter 3) for an overview. Briefly, Romani shares some innovations that took place in Indic languages between the Old Indo-Aryan (OIA) period and the Middle Indo-Aryan (MIA) period, such as the loss of consonant clusters (OIA *ra**kt**a* > MIA *ra**tt**a*, Romani *ra**t*** Ablood") – which suggests that the ancestors of Romani speakers were in Central India before the 4th century BC. On the other hand, Romani preserves a number of OIA conservatisms, such as the preservation of some other clusters (OIA *pa**tr**a*, MIA *pa**tt**a*, Romani *pa**tr**in* Aleaf") that were lost in the Central languages in Middle Indic by the 4th century BC. These are also preserved in most of the Dardic languages of the Northwest. Further, there are a number of phonological innovations that took place in Romani that did not take place in any of the Indic languages. The change of voiced aspirates to unvoiced aspirates is a case in point, e.g. Indic *bhen* > Romani *phen* "sister": these changes suggest strongly that there is a form of early Romani spoken by one speech community, as all Romani varieties underwent this change, and none of the languages of India. These developments – partly shared and partly not shared with languages in South Asia – suggest that the ancestors of Romani speakers left Central India while some of these changes occurred, i.e. the 4th century BC.

Where did they go to when they left Central India? Some shared features of languages of Northern India suggest that the speakers of ancestral Romani lived in Northern India, with speakers of Dardic and Nuristani languages. These languages also share the preservation of some older consonant clusters (OIA *tri*, MIA *ti*, Dardic Kashmiri *tri*, Romani *trin* Athree"). In addition, one can point to a number of grammatical developments shared with Kashmiri spoken in North India. Both have oblique noun endings in -s, Kashmiri/Dardic and Romani show similar innovations in the development of the participle, that are probably connected to the development of ergativity in Central Indic, which spread from the south, but not in Romani. The causative *-ar* may also be a development connected with the Dardic area (Matras 2002:43).

5.9. Iranian languages

The borrowings from Iranian languages partly show older forms, and therefore one is forced to date the presence of the ancestors of the Roma rather early (Hancock 1995, Tcherenkov & Laederich 2004:20). The contact with Iranian speakers must have been fairly intensive and of considerable duration because there is also significant grammatical influence. The tense system of Romani was reorganized through the addition a form of the copula to the past tense in order to indicate remote past on an Iranian model (Matras 2002:154). Possibly the use of interrogatives for conjunctions and the reduction of the infinitive and a few other developments were also modelled on Iranian languages (Matras 2002:154, 196).

5.10. Written language sources: Armenia

There are a few dozen Armenian loans in Romani (Boretzky 1995). These allow us to date a presence in Armenia. Armenian underwent a very unusual sound change from /l/ to a voiced velar fricative /ɣ/ (a sound close to the French <r>), which has been dated as started in the eighth century and completed in the tenth century. Three of the Romani borrowings from Armenian are *momeli* "candle", *phol* "gold and *thalik* "coat", which reflect an older pronunciation of Armenian *momełen* "wax", *phoł* "money, treasure" and *thałikc* "coat" (Tcherenkov & Laederich 2004:24). This indicates contact with Armenian before the 11th century.

5.11. Greek

Greek had a lasting influence on Romani, not only on the lexicon (Grant 2003), but also on the grammatical system. There seem to be several temporal layers, including Byzantine Greek and Medieval Greek, in the lexical component, whereas the grammatical impact is also considerable. It is very well possible that some of the Greek influence can be related to Greek as spoken in Anatolia before the spread of Turkish. Among others, the system of adopting (European) loanwords was copied from Greek (Bakker 1997).

5.12. Periodization

On the basis of linguistic data and nothing but linguistic data, it appears to be possible to reconstruct a fairly detailed **migration route** for the ancestors of the Gypsies. Not only can a rough route be established, it is also possible to put tentative dates to these events. There are competing theories with regards to dating and not so much with regards to the route itself. The earliest of the migration theories was probably Turner (1926, 1927), in discussion with Sampson (1926), who added much more detail to the picture. Kaufman's theory in Campbell (1998) comes close to the one presented here.

Even though some details remain unclear, we can try and integrate Matras (2002) and Tcherenkov & Laed-

erich (2004), whose base their studies on the earlier sources, and present the following time schedule:

Period I: presence in Central India, before the 5th/6th century BC (simplification of consonant clusters started in Central India in the 4th century BC, but did not effect Romani).

Period II: presence in Northern India (shared conservatisms with Dardic languages, such as preservation of two sibilants, in the Middle Indic period, before 1000 AD).

Period III: presence in Persia before Islamization, at the latest in the 7th century (absence of Arabic loans, present in local languages before 9th century; some Persian loanwords point to pre-Islamic forms).

Period IV: presence in Armenia before the end of the 10th century (sound change in Armenian of /ɬ/ to /χ/ in Armenia was completed by the 1oth century, but Romani has /l/ for Armenian /ɬ/).

Period V: presence in Greek-speaking area (Anatolia?) of the Byzantine Empire between 950–1200 CE (hundreds of Greek loans, many from medieval Greek).

Period VI: migration into Europe and split-up in groups, after 1200 AD (no shared pool of loanwords; first unambiguous historical sources).

These dates are on the whole earlier than what was a form of consensus for the last decade or so. Usually, a date of 1000 CE was proposed for outmigration, linked to a period of upheaval and unrest in India. Newer data and better knowledge of the languages involved present us with better possibilities, and the scenario above seems reasonable and in keeping with the known facts.

6. Conclusions

Purely linguistic evidence enables us to date the emigration of the ancestors of the Roma on the basis of documented sound changes in India. Furthermore, we can put tentative dates to the different periods spent in the areas where these languages were spoken, on the sole basis of such facts as the absence of loans from Arabic and the fact that some Persian borrowings must be quite old, and the absence of a datable sound change that took place in Armenian but not in Armenian loans in Romani. Also, on the basis of the preservation of significant numbers of loanwords, in addition to grammatical influences from Dardic-Nuristani languages, Persian and especially Greek, we have to assume that the ancestors of the Roma spent extended periods in those areas, since such developments usually take time. Finally, the fact that the shared component from Asian languages in Romani is so homogenous, forces us to accept that the people who became known as Gypsies in Europe, must have travelled **in one group** from India to South Eastern Balkans, from where they split.

As shown, linguistic changes can shed light upon developments in the past, where archaeology cannot help. Linguistic data can in fact be more reliable than historical documents since they don't rely on any one author's truthfulness. Therefore, linguistic facts can be seen as symbolic archaeological artefacts, and should be of great interest to not only to linguists but to archaeologists as well.

References

Bakker, P. 1997. Athematic morphology in Romani: The borrowing of a borrowing pattern. In Matras, Y., Bakker P. & Kyuchukov, H. (eds.). *The Typology and Dialectology of Romani,* pp. 1–21. John Benjamins, Amsterdam.

Bakker, P. 2009. Genetic roots of the Roms. In Kyuchukov H. (ed.). *New Studies in Romology*, pp. 17–45. Wini 1837, Sofia.

Bakker, P. In press. Romani genetic linguistics and genetics: results, prospects and problems. To appear in Romani Studies, Liverpool.

Bakker, P., Hübschmannová, M., Kalinin, V., Kenrick, D., Kyuchukov, H., Matras, Y. & Soravia. G. 2000. *What is the Romani language?* University of Hertfordshire Press, Hatfield.

Boretzky, N. 1995. Armenisches im Zigeunerischen (Romani und Lomavren). *Indogermanische Forschungen 100*, 137–155.

Bouwer, S., Angelicheva, D., Chandler, D., Seeman, P., Tournev, I. & Kalaydjieva, L. 2007. Carrier rates of the ancestral Indian W24X mutation in GJB2 in the general Gypsy population and individual subisolates. *Genetic Testing 11, 4*, 455–8.

Campbell, L. 1998. *Historical Linguistics: An Introduction.* Edinburgh University Press, Edinburgh.

Elšík, V. 2009. Loanwords in Selice Romani, an Indo-Aryan language of Slovakia. In Haspelmath, M. & Tadmor, U. (eds.). *Loanwords in the World's Languages: A Comparative Handbook*, pp 260–303. Mouton De Gruyter, Berlin.

Elšík, V. 2009 n.d. Selice Romani Vocabulary. http://wold.livingsources.org/vocabulary/description/9, http://wold.livingsources.org/vocabulary/9 (accessed 2009).

Fraser, A. 1994. *The Gypsies*. Blackwell, Oxford.

Gilsenbach, R. 1998. *Weltchronik der Zigeuner, Teil I: Von den Anfaengen bis 1599.* Peter Lang, Bern.

Grant, A.P. 2003. Where East and West meet. Observations on a list of Greek loans in European Romani. In Grant A.P. (ed.). *Papers in Contact Linguistics,* pp. 27–69. Special issue of Bradford Studies in Language, Culture and Society 6.

Gray, R.D. & Atkinson, Q.D. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature 426, 27,* 435–439.

Gray, R.D. & Jordan, F.M. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature 405,* 1052–1055.

Greenhill, S.J., Atkinson, Q.D., Meade, A. & Gray, R.D. 2010. The shape and tempo of language evolution. *Proc. R. Soc. B 2010 277*, 2443–2450.

Greenhill, S.J., Drummond, A.J. & Gray, R.D. 2010. How Accurate and Robust Are the Phylogenetic Estimates of Austronesian Language Relationships? *PLoS One 5, 3*, e9573.

Gresham, D., Morar, B., Underhill, P., Passarino, G., Lin, A.A., Wise, C., Angelicheva, D., Calafell, F., Oefner, P.J., Shen, P., Tournev, I., de Pablo, R., Kuchinskas, V., Perez-Lezaun, A., Marushiakova, E., Popov, V. & Kalaydjieva, L. 2001. Origins and divergence of the Roma (Gypsies). *American Journal of Human Genetics 69*, 1314–31.

Hancock, I.F. 1993. Lo studente Ungherese Valyi Istvan e le origini indiane della lingua romani [The Hungarian student Valyi Istvan and the Indian origin of the Romani language]. *Lacio Drom* 29, 5, 21–23.

Hancock, I.F. 1995. On the migration and affiliation of the ?mba: Iranian words in Rom, Lom and Dom Gypsy. In Matras, Y. (ed.). *Romani in contact. The history, structure and sociology of a language,* pp. 25–51. John Benjamins, Amsterdam.

Haspelmath, M. & Tadmor, U. 2009. *Loanwords in the World's Languages: A Comparative Handbook*. Mouton De Gruyter, Berlin.

Heine, B. & Kuteva, T. 2005. *Language Contact and Grammatical Changes*. Cambridge University Press, Cambridge.

Kalaydjieva, L., Gresham, D. & Calafell, F. 2001. Genetic studies of the Roma (Gypsies): a review. *BMC Medical Genetics 2:5*.

Kalaydjieva, L., Morar, B. Chaix, R. & Tang, H. 2005. A newly discovered founder population: the Roma/Gypsies. *Bio-Essays 27*, 1084–1094.

Ioviță, R. & Schurr, T. 2004. Reconstructing the origins and migrations of diasporic populations: the case of the European Gypsies. *American Anthropologist 106, 2,* 267–281.

Liégeois, J.P. 2008. *Roma in Europe.* Council of Europe, Strasbourg.

Masica, C. 1991. *The Indo-Aryan languages*. Cambridge University Press, Cambridge.

Matras, Y. 1999. Johann Rüdiger and the study of Romani in 18th-century Germany. *Journal of the Gypsy Lore Society, fifth series, 9*, 89–116.

Matras, Y. 2002. *Romani: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Morar. B., Gresham, D., Angelicheva, D., Tournev, I., Gooding, R., Guergueltcheva, V., Schmidt, C., Abicht, A., Lochmuller, H., Tordai, A., Kalmar, L., Nagy, M., Karcagi, V., Jeanpierre, M., Herczegfalvi, A., Beeson, D., Venkataraman, V., Warwick, C.K., Reeve, J., de Pablo, R., Kucinskas V. & Kalaydjieva, L. 2004. Mutation history of the Roma/Gypsies. *Am J Hum Genet 75, 4,* 596–609.

Navarro, C. & Teijeira, S. 2003. Neuromuscular disorders in the Gypsy ethnic group. A short review. *Acta Myol 22, 1,* 11–4.

Pischel, R. 1883. Die Heimath der Zigeuner. *Deutsche Rundschau 36*, 353–375.

Renfrew, C., McMahon, A. & Trask, L. 2000. *Time Depth in Historical Linguistics*. 2 Vols. McDonald Institute for Archaeological Research, Cambridge.

Sampson, J. 1911. Jacob Bryant: being an analysis of his AngloRomani vocabulary, with a discussion of the place and date of collection and an attempt to show that Bryant, not Rüdiger, was the earliest discoverer of the Indian origin of the Gypsies. *Journal of the Gypsy Lore Society New Series*, *4,* 162–94.

Sampson, J. 1927. Notes on Professor R.L. Turner's "The position of Romani in IndoAryan". *Journal of the Gypsy Lore Society, Third Series*, *6,* 57–68.

Tcherenkov, L. & Laederich, S. 2004. *The Rroma*. 2 Volumes. Schwabe, Basel.

Thomas, P., Kalaydjieva, L., Youl, B., Rogers, T., Angelicheva, D., King, R., Guergueltcheva, V., Colomer, J., Lupu, C., Corches, A., Popa, G., Merlini, L., Shmarov, A., Muddle, J., Nourallah, M. & Tournev, I. 2001. Hereditary motor and sensory neuropathy - Russe: new autosomal recessive neuropathy in Balkan. Gypsies. *Ann. Neurol. 50, 4*, 452 –457.

Turner, R.L. 1926. The position of Romani in IndoAryan. *Journal of the Gypsy Lore Society, Third Series, 5*, 145–189.

Turner, R.L. 1927. The position of Romani in IndoAryan: a reply to Dr. J. Sampson. *Journal of the Gypsy Lore Society, Third Series*, *6*, 129–138.

Willems, W. 1997. *In Search of the True Gypsy. From Enlightenment to Final Solution*. Routledge, London.