# Roma: reconstruction migrations with linguistic and genetic data

Peter Bakker

Linguistics, Aarhus University, DK

Workshop on Migration, Porquerolles, September 2007

Linpb AT hum.au.dk

- With the more fancy slides made by Yaron Matras

- I.  Who are the Roma?
- 2.  How can linguists use their data to make historical inferences?
- 3. Romani dialects/varieties
- 4.  The migration of the Roma from India based on linguistic data
- 5.  Support from genetics

# Who are the Roma

- Alias Gypsies, Tsiganes, Zigeuner, Roma and Sinti ......
- Ca. 6 million people mostly in (Eastern) Europe
- Popular image: nomads
- Reality: settled people
- (but they came from elsewhere, perhaps as long as 4-5 centuries ago)
- Strong in-group attitudes
- Strong (extended) family bonds

# Nomads versus Gypsies

- ”Gypsies” as used by outsiders: nomadic people
- Term used by insiders: a specific ethnic group, or an amalgam of ethnic (sub)-groups

- Extremes of definitions:
  - any nomadic group or person (”Gypsy scholar”)
  - Those who speak a language called Romanes, and who are born into specific families

# ”Gypsies”: out of (Little) Egypt?

- Insiders’ labels:
  - **Rom** ”human, man” < *Dom* ”man, caste/ethnic group of smiths and musicians”; also Romano, Romanichel.
  - **Sinti** < ??? [not cognate with *Sindh* or *Hindi*]
  - **Manuš <** Romani *manuš* ”person”
  - **Calo/Kaalo** < Romani *kalo* ”black”
- Outsiders’ labels:
  - **Egyptian** > Gypsy, Gitane, Gitano, Ijito, Jifti, etc.
  - ? **Atsingan** > çingene, Tsigane, Zigeuner, Sigøjner, etc
  - **Tatar** > Tattare, Tatere (North Germany and up)
- Language name: always ”Romanes”
- Language is also one and the same

# Names of subgroups

- Based on ***country/region, e.g.***

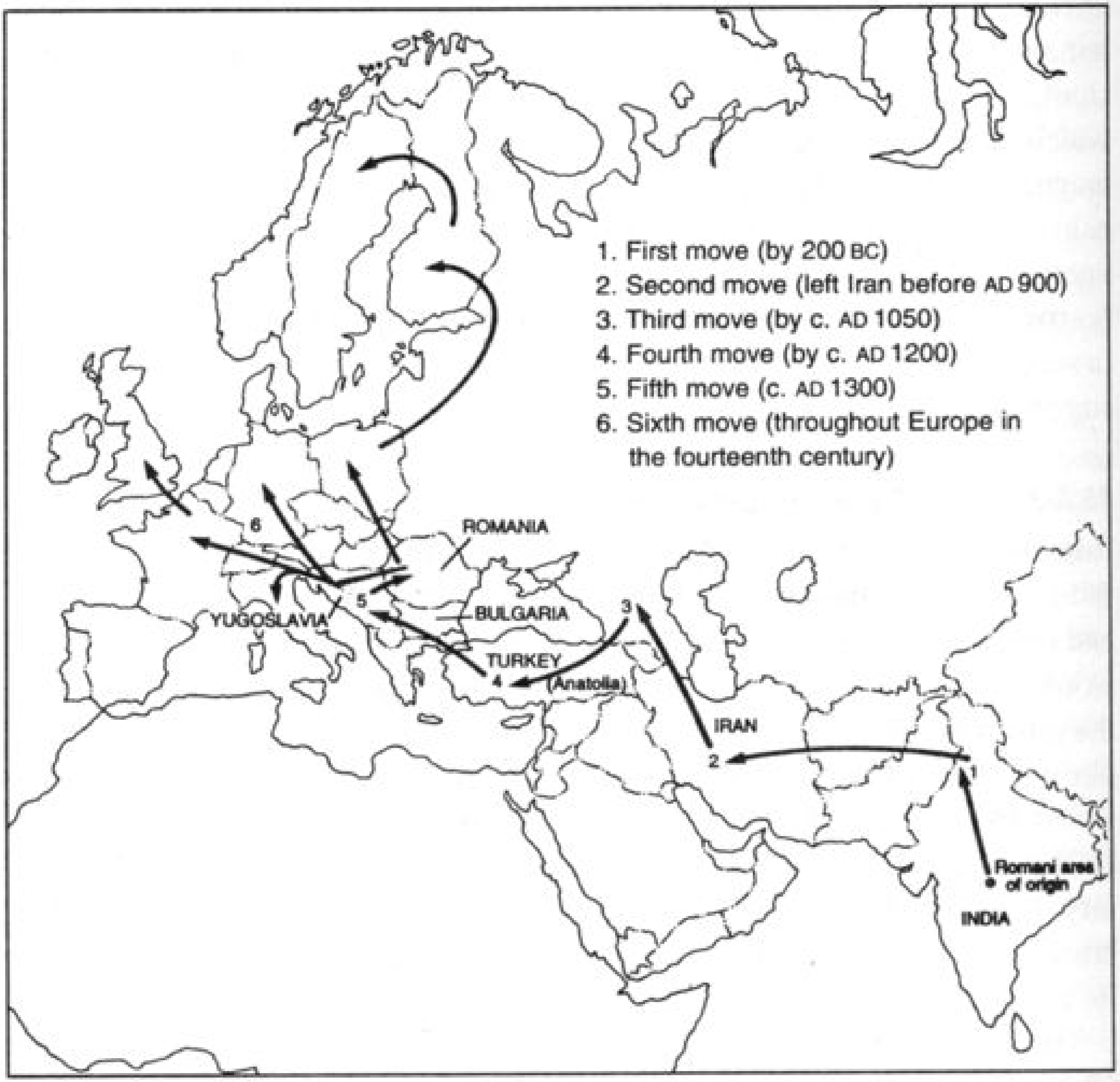
| | |
|---|---|
| Olah/Vlah | ”Roumania” (Wallachia) |
| Serbika | ”Serbia” |
| Rom-ungro | ”Hungary” |

- Based on (former) ***occupation. e.g.***

| | | |
|---|---|---|
| Lovari | < Hungarian | *lo* ”horse” |
| Kalderash | < Roumanian | *câldâr* ”kettle” |
| Kalaydži | < Turkish | *kalayci* ”tinsmith” |

# Language: Indic

- Indic connection was unknown until ca. 1780
- Link with Indic language discovered more or less simultaneously in:
  - Britain (Bryant),
  - Hungary/Netherlands (Vali),
  - Germany (Rüdiger),
  - Russia (Simon Peter Pallas/ Catherine the Great)
- ”We come from Little Egypt”: Anatolia?


1. First move (by 200 BC)
2. Second move (left Iran before AD 900)
3. Third move (by c. AD 1050)
4. Fourth move (by c. AD 1200)
5. Fifth move (c. AD 1300)
6. Sixth move (throughout Europe in the fourteenth century)
ROMANIA
6
5
YUGOSLAVIA
BULGARIA
3
TURKEY
4
(Anatolia)
IRAN
2
1
Romani area of origin
INDIA

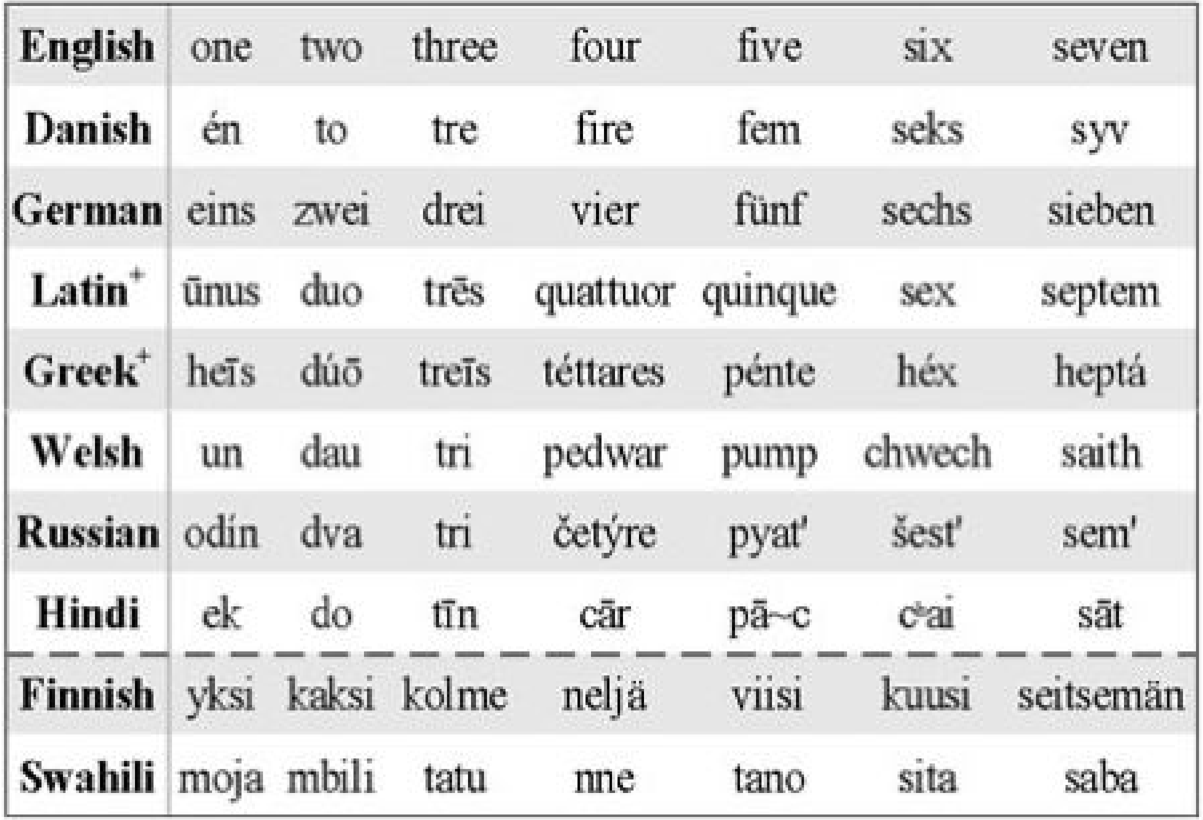
MAP 15.3: Romani (Gypsy) migrations (based on Kaufman 1973)

# Historical linguistics

How do linguists decide that two languages are ”genetically related”? I.e. a common origin?

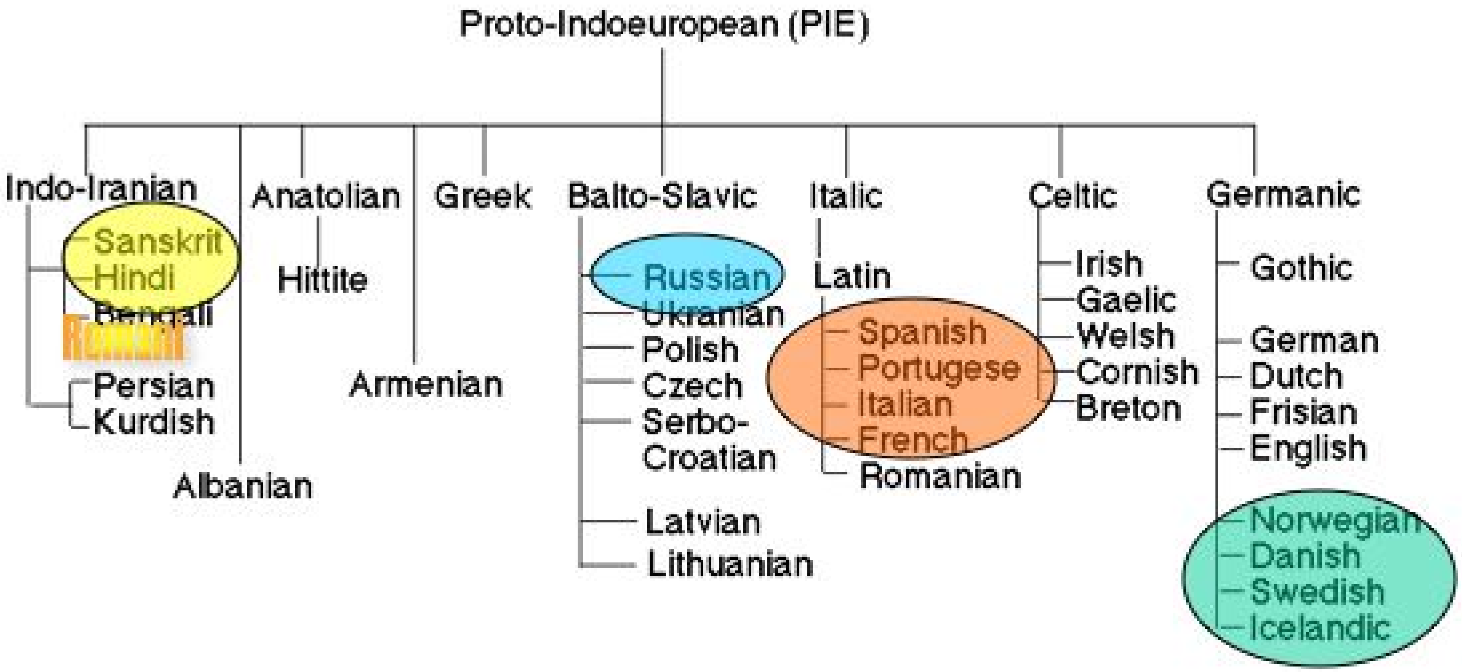
- Similarities in common **vocabulary** (body parts, numerals, kinship, weather, etc.): *form and meaning*
- Similarities in **grammatical** elements (pronouns, verbal and nominal endings, etc.) *form, meaning*
- Sound changes in these words must be **regular.**

# Comparative historical linguistics



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **English** | one | two | three | four | five | six | seven |
| **Danish** | én | to | tre | fire | fem | seks | syv |
| **German** | eins | zwei | drei | vier | fünf | sechs | sieben |
| **Latin**$^{+}$ | ūnus | duo | trēs | quattuor | quinque | sex | septem |
| **Greek**$^{+}$ | heĩs | dúō | treĩs | téttares | pénte | héx | heptá |
| **Welsh** | un | dau | tri | pedwar | pump | chwech | saith |
| **Russian** | odín | dva | tri | četýre | pyat' | šest' | sem' |
| **Hindi** | ek | do | tīn | cār | pā~c | c$^{h}$ai | sāt |
| **Finnish** | yksi | kaksi | kolme | neljä | viisi | kuusi | seitsemän |
| **Swahili** | moja | mbili | tatu | nne | tano | sita | saba |

# A Simplified Family Tree of Indoeuropean



Proto-Indoeuropean (PIE)

Indo-Iranian
Sanskrit
Hindi
Bengali
Romani
Persian
Kurdish

Albanian

Anatolian
Hittite

Armenian

Greek

Balto-Slavic
Russian
Ukranian
Polish
Czech
Serbo-Croatian
Latvian
Lithuanian

Italic
Latin
Spanish
Portugese
Italian
French
Romanian

Celtic
Irish
Gaelic
Welsh
Cornish
Breton

Germanic
Gothic
German
Dutch
Frisian
English
Norwegian
Danish
Swedish
Icelandic

# How not to prove I (after Trask)

| | *A* | *B* | *C* | *D* | *E* |
|---|---|---|---|---|---|
| *News* | Xabar | Xabar | Haber | Habari | khabar |
| *Time* | Waqt | Vaqt | Vakit | Wakati | Waktu |
| *Book* | Kitâb | Kitâb | Kitap | Kitabu | kitab |
| *Service* | Xidmat | Xidmatgari | Hizmet | Huduma | khidmat |
| *beggar* | Faqir | Faqir | Fakir | Fakiri | fakir |

# How not to prove I: borrowings!

| | *Arabic* | *Urdu* | *Turkish* | *Swahili* | *Malay* |
|---|---|---|---|---|---|
| *News* | Xabar | Xabar | Haber | Habari | khabar |
| *Time* | Waqt | Vaqt | Vakit | Wakati | Waktu |
| *Book* | Kitâb | Kitâb | Kitap | Kitabu | kitab |
| *Service* | Xidmat | Xidmat gari | Hizmet | Huduma | khidmat |
| *beggar* | Faqir | Faqir | Fakir | Fakiri | fakir |

# How not to prove II (after Trask)

| | |
|---|---|
| Aeto ”eagle” | Aetos ”eagle” |
| Noonoo ”thought” | Nous ”thought” |
| Manao ”think” | Manthano ”learn” |
| Mele ”sing” | Melos ”melody” |
| Lahui ”people” | Laos ”people” |
| Meli ”honey” | Meli ”honey” |
| Kau ”summer” | Kauma ”heat” |

# How not to prove I: coincidence!

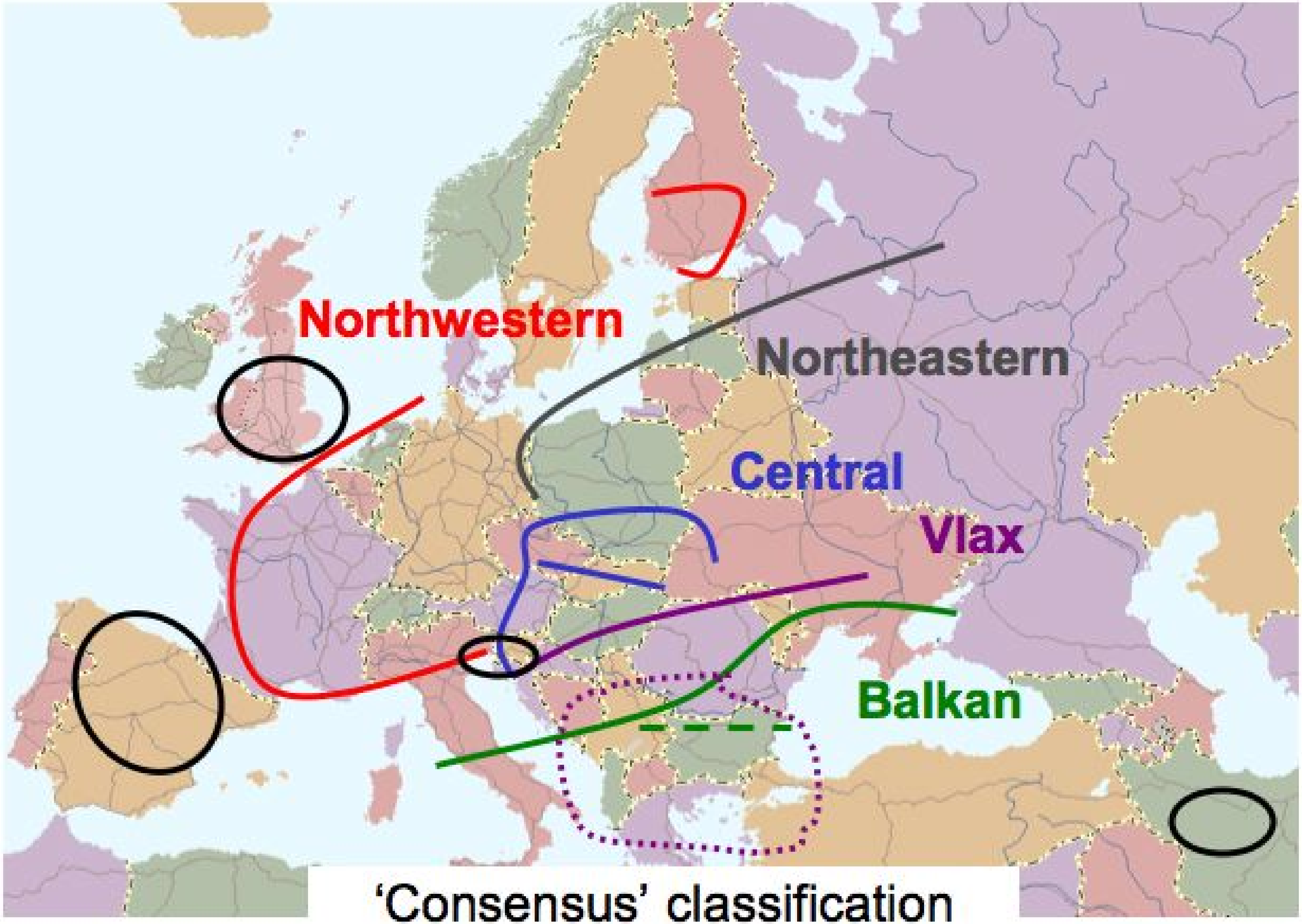
| | |
|---|---|
| Aeto ”eagle” | Aetos ”eagle” |
| Noonoo ”thought” | Nous ”thought” |
| Manao ”think” | Manthano ”learn” |
| Mele ”sing” | Melos ”melody” |
| Lahui ”people” | Laos ”people” |
| Meli ”honey” | Meli ”honey” |
| Kau ”summer” | Kauma ”heat” |
| **Hawaiian** | **Ancient Greek** |

# Romani varieties

- Linguists classify languages into **families**
- Linguist classify dialects into **groupings**
  - On the basis of shared linguistic features (common words, common structures)
- (The labels/names used by Roma are not always the same as the ones used by linguists)

# Four Romani dialect groups:

- **Vlax,** (Rumania, and from there all over
  *North + South* Europe; West Europe in 1800s)
- **Balkan,** (Kosovo, Macedonia, Turkey,
  *N + South I+II* Bulgaria, Serbia, Iran)
- **Central** *N+S* (Hungary and neighbouring
  countries)
- **North**-conglomerate (*East+West*)
  (Western Europe,
  from South Italy to Finland)



‘Consensus’ classification

# Romani in Europe

- Everybody above age 6 speaks at least one other language: universal bi/multilingualism
- Additional language knowledge is valued

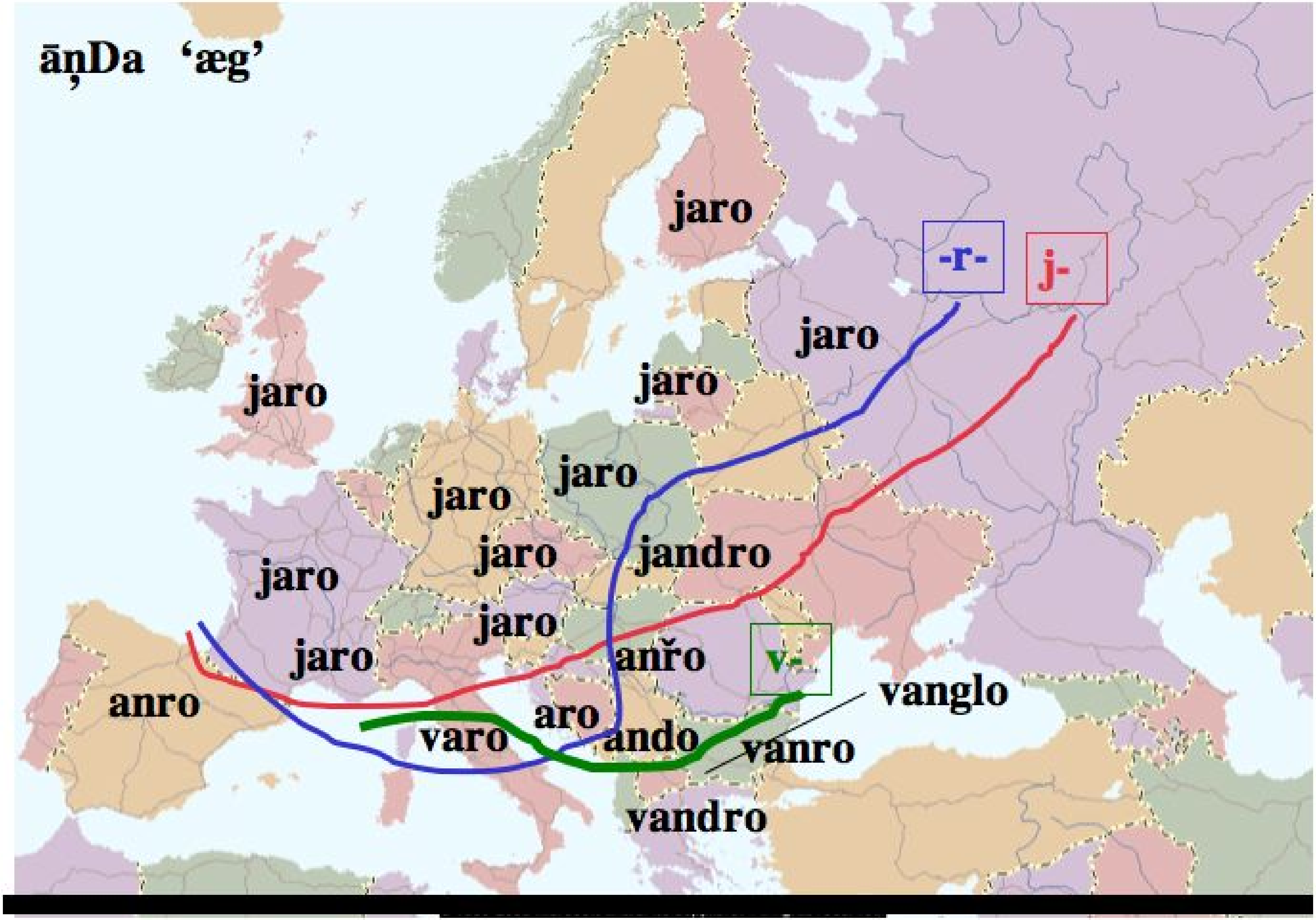
- Romani has low status language in society, high status in the family
- Low status languages borrow words and constructions from dominant languages

# Loanwords as historical sources

| | *Always loans from* | *Never loans from* |
|---|---|---|
| *Vlax* | Rumanian | S: Hungarian,<br>N: Turkish |
| *Balkan* | Turkish, Slavic | Rumanian, Hungarian |
| *Central* | Hungarian | Turkish |
| *North* | German, Northw. Slavic | Rumanian, Hungarian, Turkish |

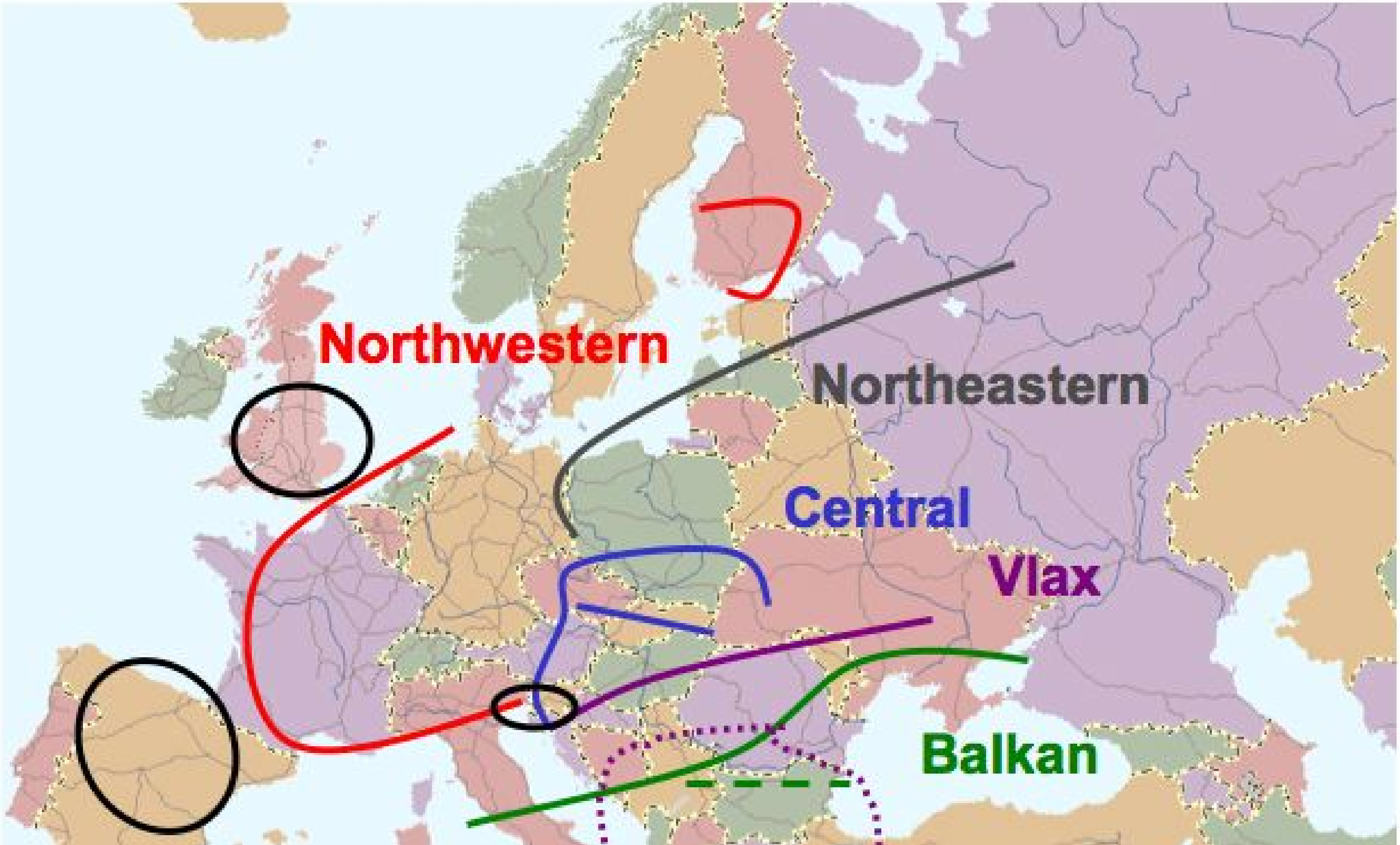
# Vertical (ancestral) and horizontal (geographic/areal)

- There are groupings based on *shared innovations*: dialect groups (inheritance, genetic, vertical)
- There are clear *geographical patterns* in sound changes (areal, horizontal)



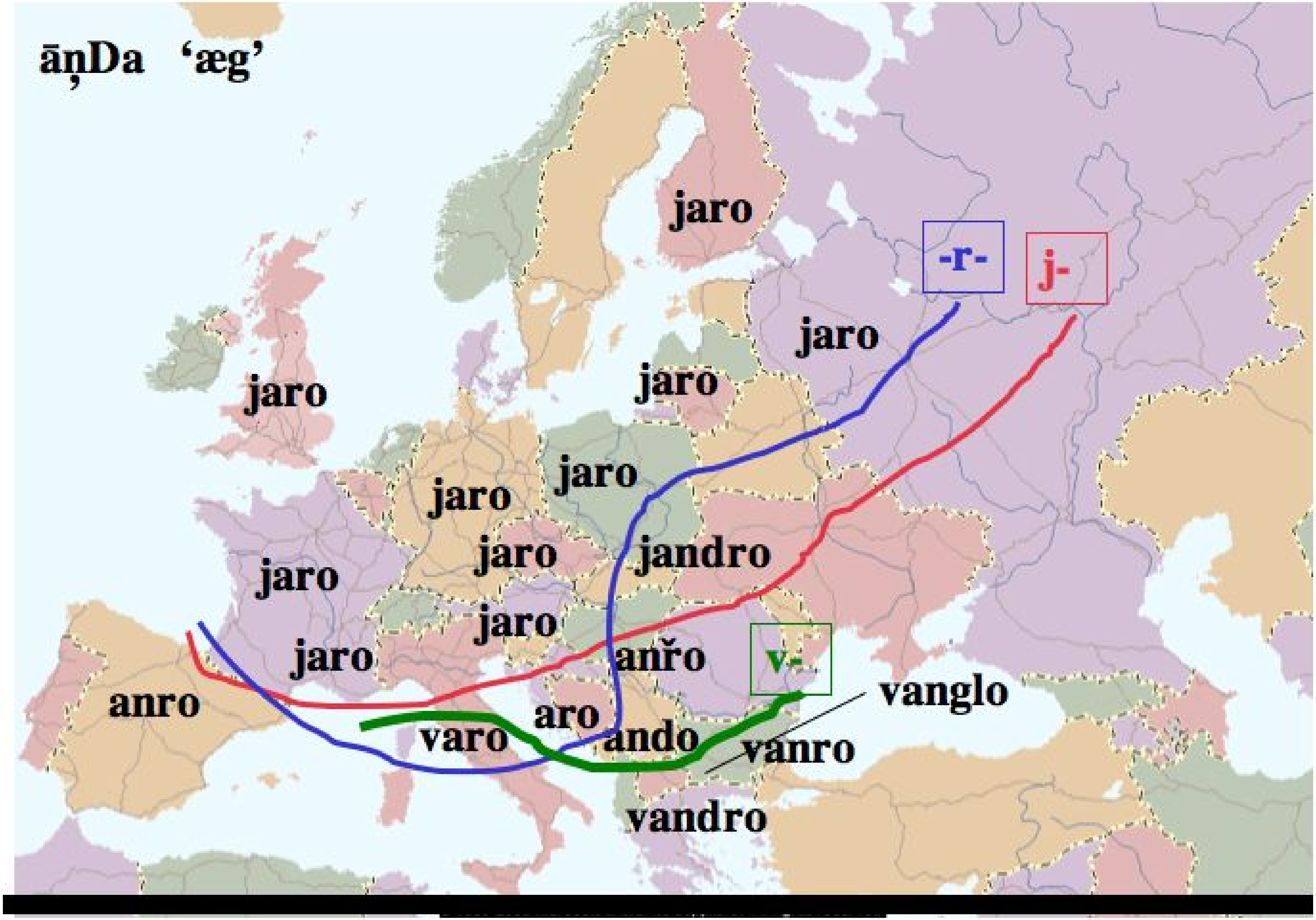
āṇDa ‘æg’

[redacted]



'Consensus' classification

āņDa ‘æg’



jaro
-r-
j-
jaro
jaro
jaro
jaro
jaro
jaro
jandro
jaro
jaro
jaro
anřo
v-
vanglo
anro
aro
varo
ando
vanro
vandro

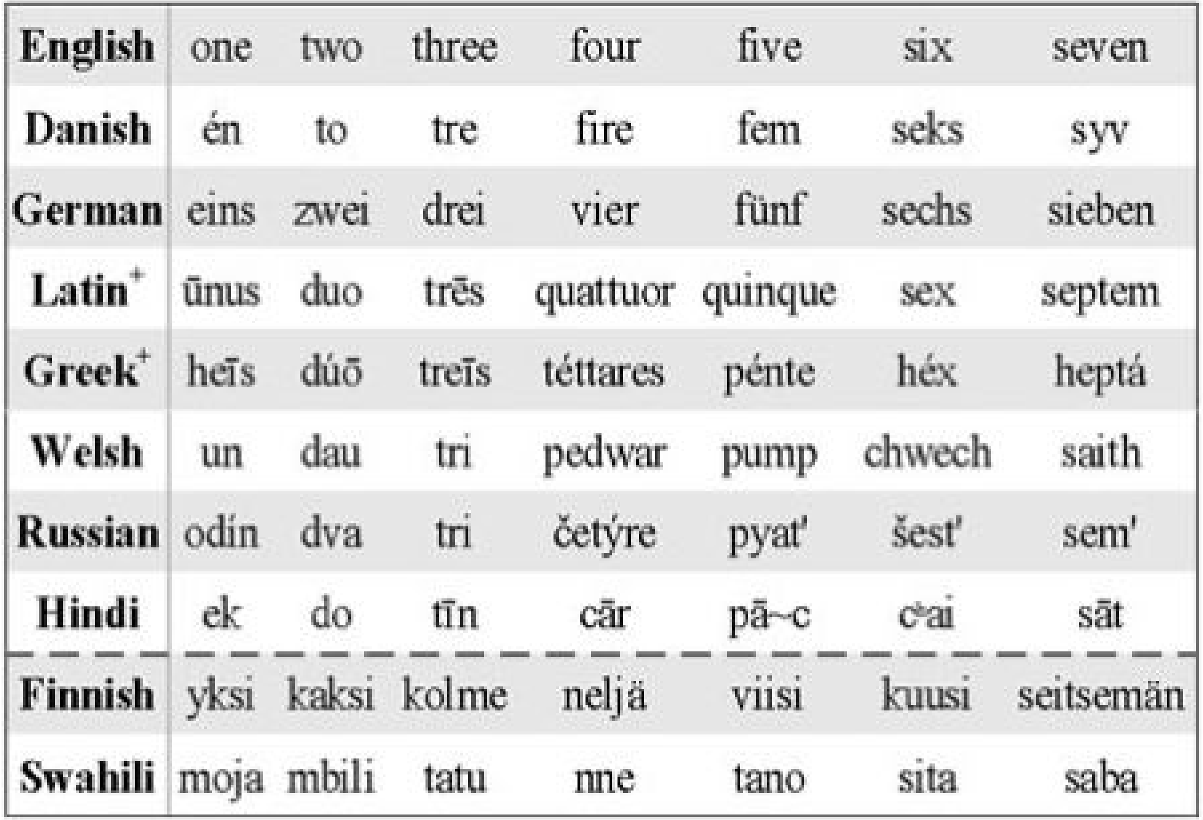
# Romani is an Indic language

- Core vocabulary is Indic: many hundreds of words
- Almost all grammatical endings are Indic
- There are many regular sound correspondences

Conclusion: Romani is an Indic language

# Romani numerals?

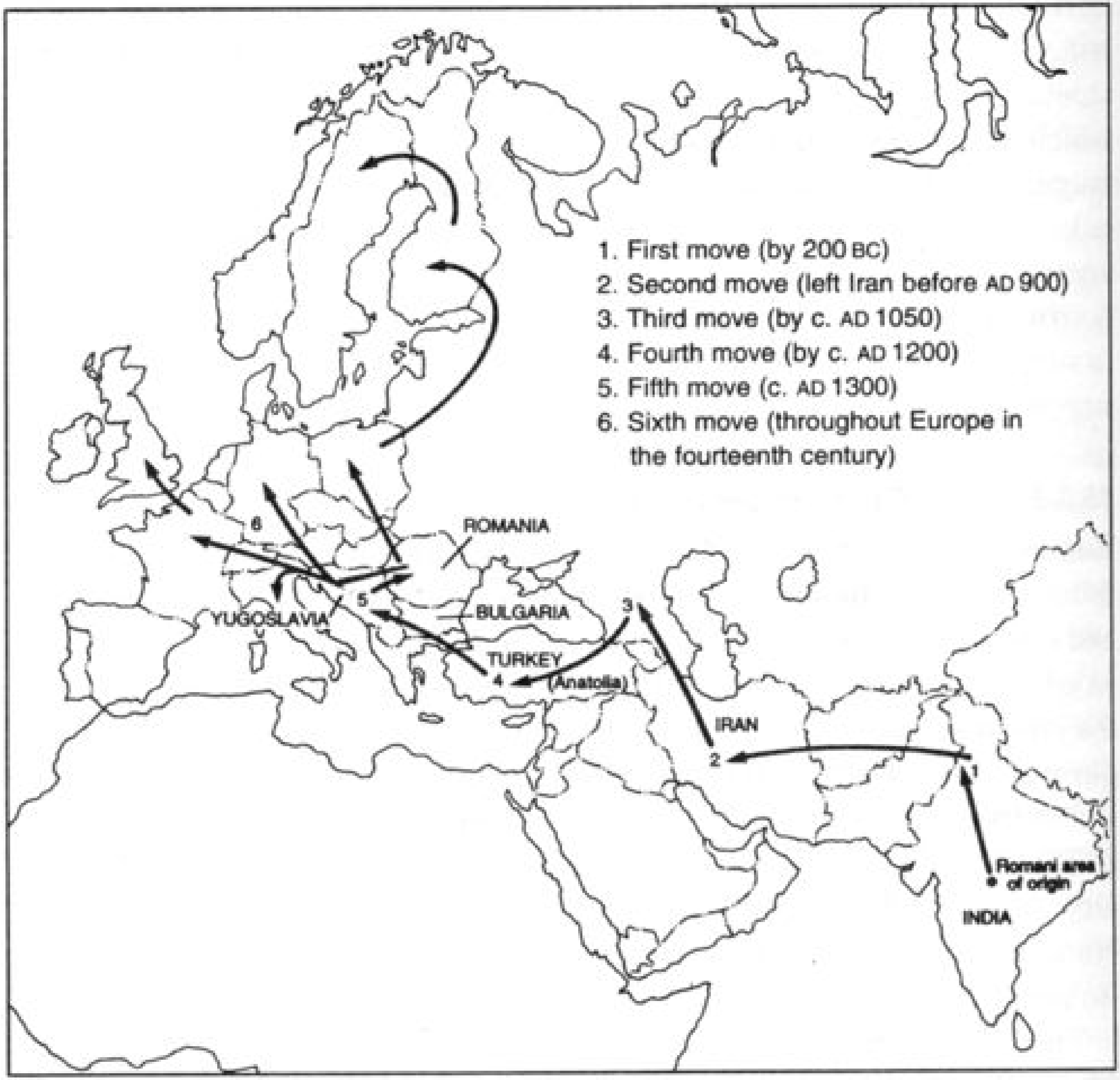
- Jekh duj trin štar pandž šov efta



| English | one | two | three | four | five | six | seven |
|---|---|---|---|---|---|---|---|
| Danish | én | to | tre | fire | fem | seks | syv |
| German | eins | zwei | drei | vier | fünf | sechs | sieben |
| Latin[+] | ūnus | duo | trēs | quattuor | quinque | sex | septem |
| Greek[+] | heîs | dúō | treîs | téttares | pénte | héx | heptá |
| Welsh | un | dau | tri | pedwar | pump | chwech | saith |
| Russian | odín | dva | tri | četýre | pyat' | šest' | sem' |
| Hindi | ek | do | tīn | cār | pā~c | cʰai | sāt |
| Finnish | yksi | kaksi | kolme | neljä | viisi | kuusi | seitsemän |
| Swahili | moja | mbili | tatu | nne | tano | sita | saba |

# How did the Roma travel from India to Europe?

- (the dating is controversial,
- the route hardly so)



1. First move (by 200 BC)
2. Second move (left Iran before AD 900)
3. Third move (by c. AD 1050)
4. Fourth move (by c. AD 1200)
5. Fifth move (c. AD 1300)
6. Sixth move (throughout Europe in the fourteenth century)
ROMANIA
6
5
YUGOSLAVIA
BULGARIA
3
TURKEY
4
(Anatolia)
IRAN
2
1
Romani area of origin
INDIA

MAP 15.3: Romani (Gypsy) migrations (based on Kaufman 1973)

# Summary migration

- Left India (between 300 and 500 A.D.?)
- Left Persia before 700
- Left Armenia before 900
- Stayed in Greek-speaking area (40.000 people?)
- They travelled in one group (shared loanwords)
- Arrival in Europe around 1200
- From there: spread from Balkan in smaller groups (30-200 people?)

# What evidence is there for this origin and travel route?

- Archaeological evidence: zero
- Historical: Pre-European historical documents: very, very little
- Historical documents along the route: close to zero
- Early European history (1400s): little
- Linguistic evidence: convincing
- Biological evidence (genetics): growing

# Common lexicon of Romani

- Oldest layer: Indic (Central India)
- Some words: Dardic (North India/Pakistan mountains)
- Many Iranian (Persian, Kurdish)
- Armenian (few dozen)
- Georgian (few)
- Many Greek (Anatolia?)
- South Slavic (common core?)

*Do these reflect a migration route???*

# 3500 years of documented history of language in India

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

# Grammatical layers of Romani

- Central Indic morphology
- Some Indic phonemes (aspirates, some retroflex)
- Some Old Indo-Aryan and Middle Indo Aryan conservatisms
- Some shared innovations with New Indo-Aryan
- Dardic traits (Northwest India)
- Iranian influence on verb
- Greek and Balkan influence on syntax and morphology

# Spread in Europe

- Vlax has spread in the past 200+ years through (first) South Balkan, (later) rest of Europe
- One can reconstruct some of the migrations with linguistic data (mostly because of **loanwords**)
- **One example:**
- ”North” groups have travelled all over Europe, e.g. **Scandinavian** Travellers/Kaalo (NO, SE, FI):
- Balkan (ca. 1300)– Czechia - Germany/France –

  England – Scotland (1505) – Denmark – Sweden/Norway - Finland

# Gypsy ”sightings” in the 14th and 15th centuries

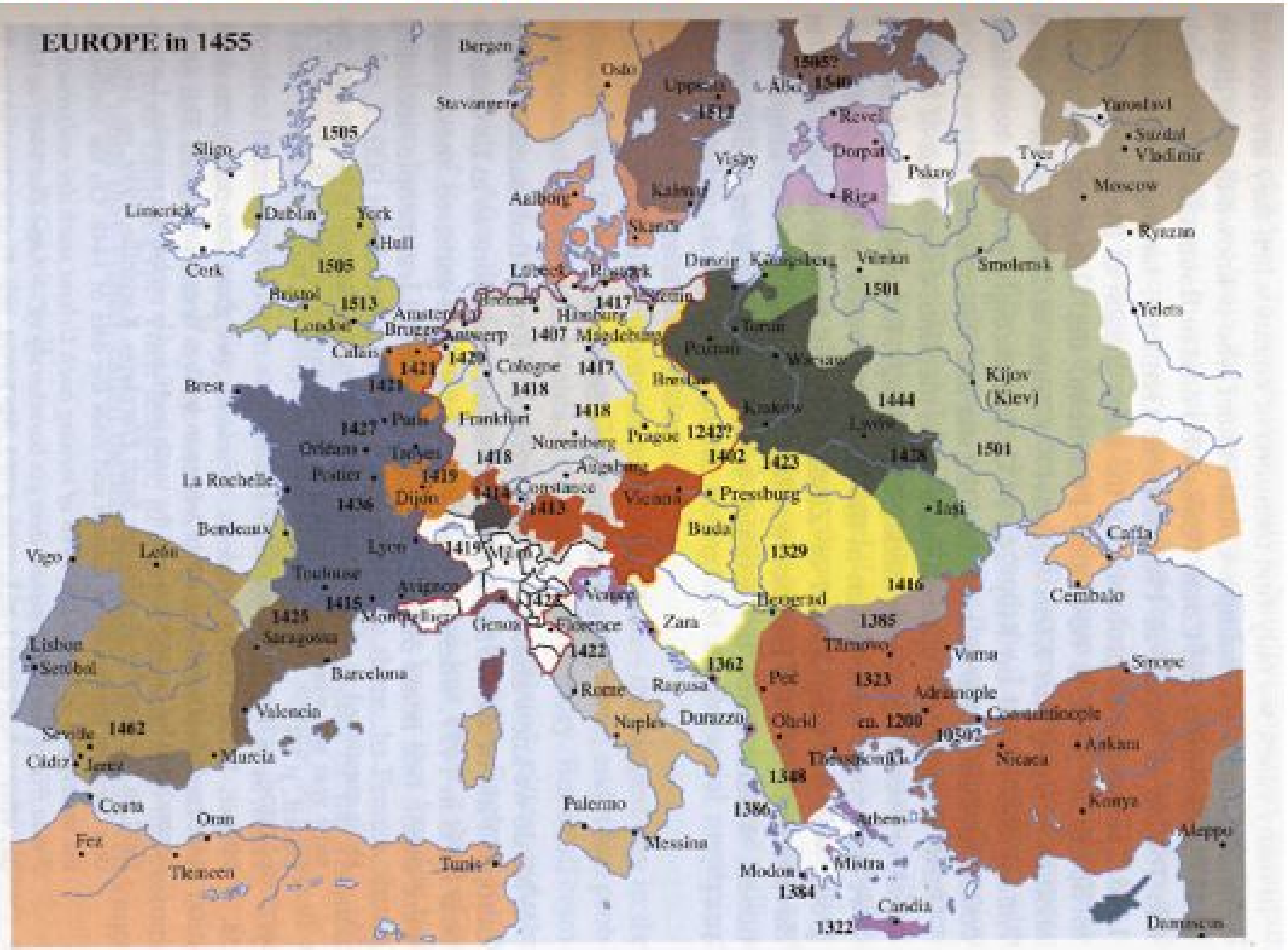


*Figure 5.1. Europe in the Early Fifteenth Century – Rroma Sightings*[35]

# Romani dialect diversification

- Diversification started between 1200-1300
- when Roma groups spread through Europe

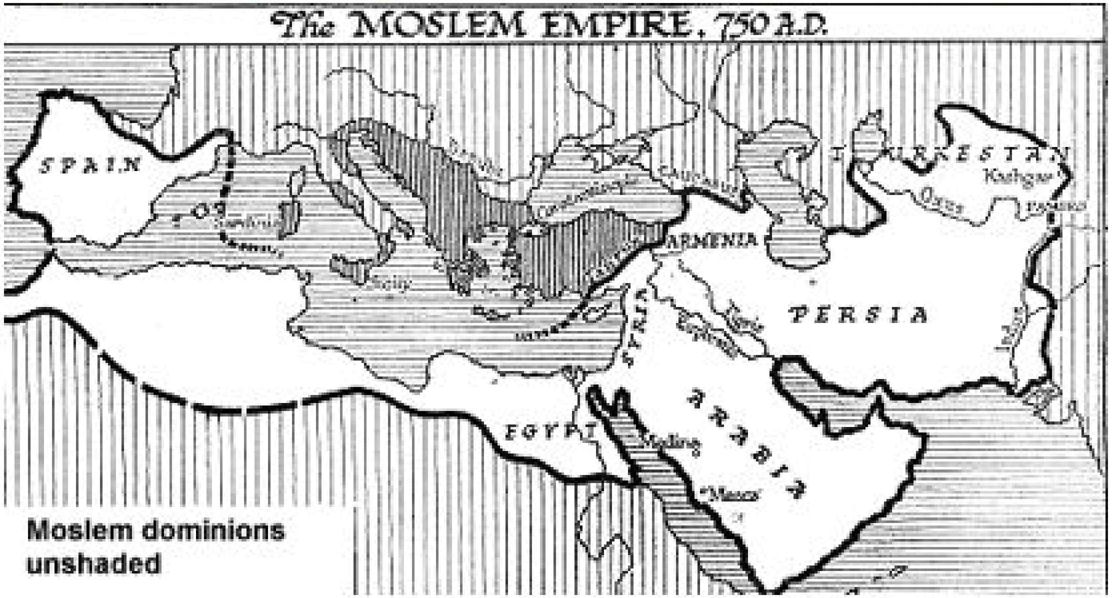
- Most groups were settled around 1550
- (some groups continued to travel seasonally for goods & services)
- (The settled Vlaxs (former slaves) became nomads! In the 1800s)

# Some problems

- 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?

# No Arabic loans?

- There are no loans from Arabic in Romani
- Strange: Arabic has spread through Asia (Iran, Pakistan) and has left borrowings



Moslem dominions unshaded

# Do the Roma come from India?

- If Romani is an Indic language, do the Roma come from India?
- (migration of **population**, or transmission of **language**?)

- Linguistics: no direct proof
- Circumstantial evidence:
  - Cultural data?
  - Physical features
  - Genetic data

# Cultural data common between ”India” and Roma

- Indian castes, Romani occupational groups: intermarriage within occupational groups
- Taboos on food: food preparation, food sharing, hygiene.
- Inherited or developed? No way to be certain

# Genetic data?

- Did (the bulk of) their ancestors come from India?
- *OR:*
- Are Roma local people who were stigmatized into becoming a new ethnic group? (Leo Lucassen)
- Who had acquired some ”Sanskrit” as used along trade routes? (Wim Willems)

# Genetic research on Roma: 3 types

- > Genetically transmitted diseases
- > Group features

  (e.g. Roma versus local populations;

  Roma versus Asian populations;

  Roma group A vs. Roma Group B)
- (Health issues, intervention)

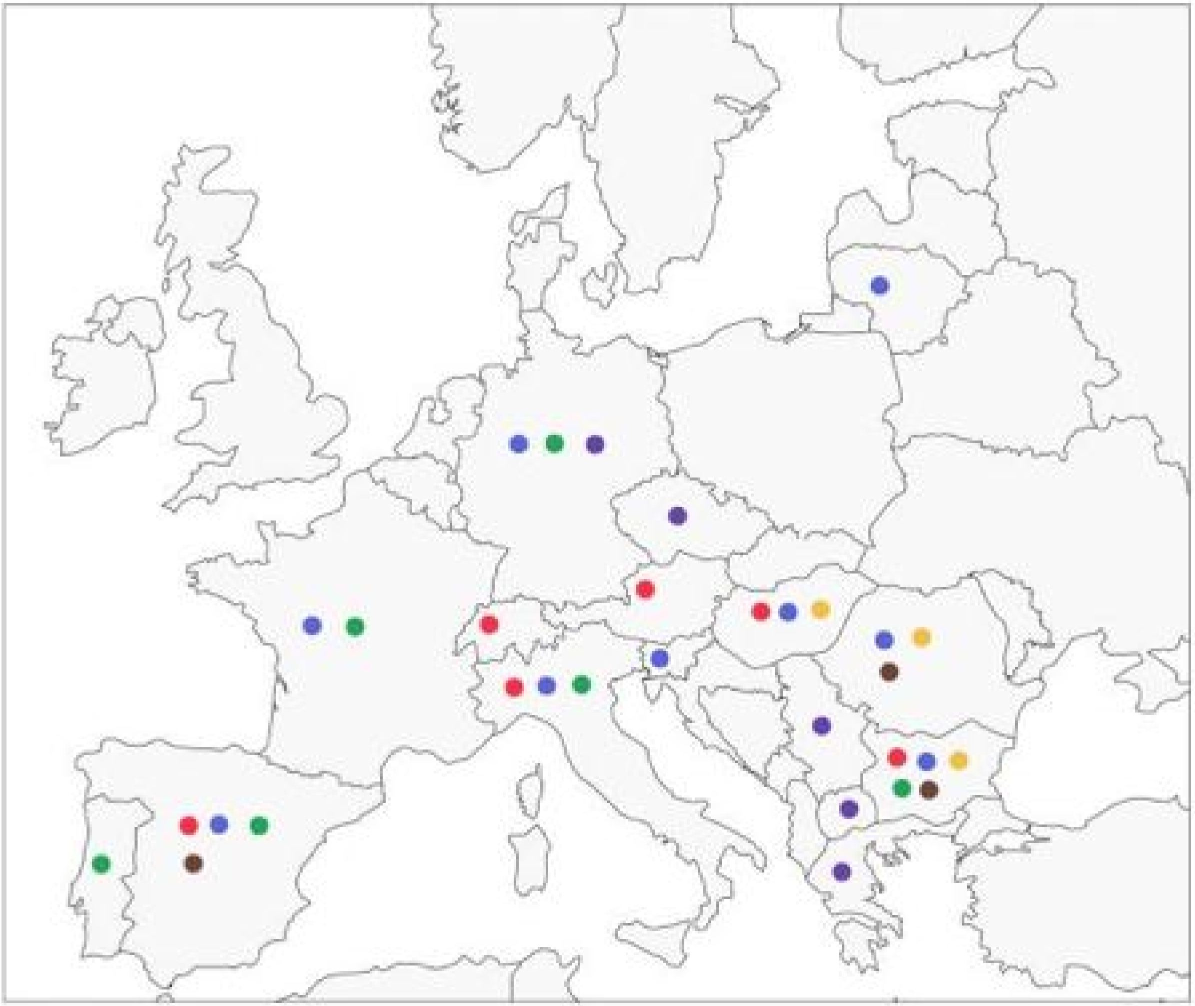
- Hundreds of papers

# Muscular Dystropy γSG (Piccolo 1996)

- Genetic muscle disease found (1996) in 7-10 Gypsy families in France, Spain and Italy.
- ”If generation is 20 years, this would indicate that the C283Y mutation in the [gamma]-SG gene is at least 1200 years old. If the genetic clock had been reset by a bottleneck around the time of wandering (fixation of one haplotype in a small population), the age of the mutation would be even older. We therefore assume that the C283Y mutation predates the commonly accepted date of migration of Gypsies out of Northern India.”

# Five rare single-gene diseases, relatively common among Roma

- CMS: several Roma groups, South Asia
- HMSNL: Roma individuals across Europe, or ”more common among the Vlax”
- CCFDN: only among ”Vlax Roma”
- LGMDC2: Western European Roma, some Balkan (not elsewhere)
- GD: across Europe, most common Vlax



- 🔴 Infantile cataracts due to galactokinase deficiency (GALK)
- 🔵 Hereditary motor and sensory neuropathy type Lom (HMSNL) 1150
- 🟡 Congenital cataracts facial dysmorphism neuropathy (CCFDN) syndrome 500
- 🟢 Limb gridle muscular dystrophy type 2C (LGMD2C) 600
- 🟣 Congenital myasthenia
- 🟤 Hereditary motor and sensory neuropathy type Russe (HMSNR)

# Dates?

- CMS: several Roma groups, South Asia:

  (800 years: 650-1025)
- HMSNL: across Europe, or ”more common Vlax” (850 years: 700-1075)
- CCFDN: only among Vlax Roma

  (500 y: 400-650)
- LGMDC2: Western European Roma, some Balkan (not elsewhere; no Vlax)

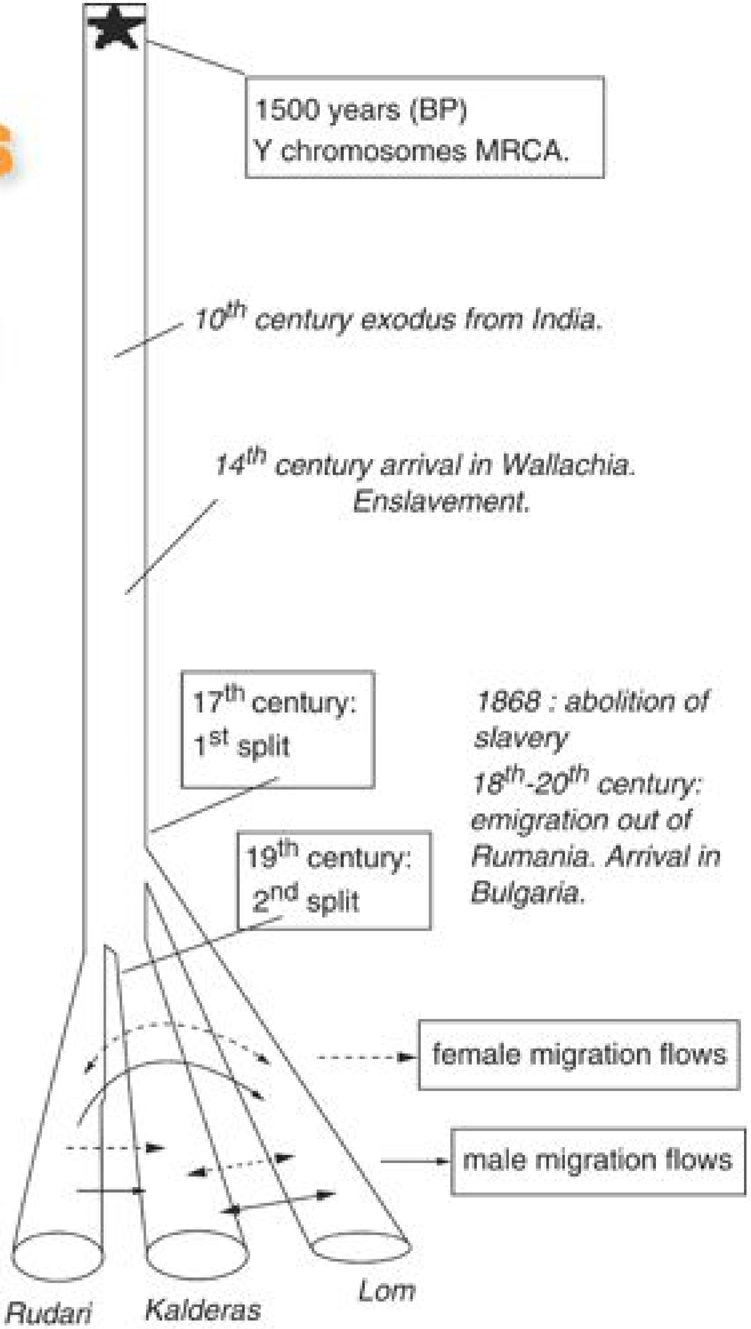
  (600 y: 525-775)

# Conclusions Morar et. al

- ”Obvious founder effect in Western European Gypsies” (603) because of ”high carrier rates” (604)
- ”The patterns observed were distinctly different in the four Gypsy groups, reflecting an independent history of random recombinations” (603)
- ”profound secondary and tertiary bottleneck events”

- ”time frame (...) is unclear”: ”founding of the Proto-Gypsy population (..) single recent founding event ca. 800-900 years ago” (upper limit). ”founded by a small group of related individuals”
- subgroups: 425-600 y. (Vlax: ~500 years ago)
- Founders of Vlax subgroups: ~100 individuals
- Limited gene flow
- ”history of endogamy is as old as Gypsy groups” (607)
- ”exodus, diaspora, and subsequent fragmentation into small, geographically dispersed and isolated communities” (607)

# females and males (three adjacent groups)



1500 years (BP)
Y chromosomes MRCA.

*10th century exodus from India.*

*14th century arrival in Wallachia. Enslavement.*

17th century:
1st split

*1868 : abolition of slavery*

*18th-20th century: emigration out of Rumania. Arrival in Bulgaria.*

19th century:
2nd split

female migration flows

male migration flows

# General conclusions genetics

- Roma are genetically closer to Indians than to European populations
- Roma are heterogenous between countries
- Internal diversity of the Roma: genetically far more heterogeneous than autochthonous European populations.
- Single locus comparisons have resulted in controversy, with some pointing to close genetic affinity between Roma and Indians, and others indicating that the Roma are indistinguishable from Europeans.

# Conclusion

- Detailed **migration route** reconstructed for Gypsies

  in the absence of archaeological or historical evidence
- Gypsies must have travelled **in one group** from India to the Balkans (> 30.000 people?)

- *Linguistic evidence:*
  - **Dating** of outmigration on the basis of documented sound changes in India
  - And on the basis of absence of loans from Arabic and shared ones from Armenian
  - **Locations** of extended stays on the basis of quantity of loans (Iranian, Greek)
  - Grammatical influence (Dardic, Iranian, Greek) tends to take time: **duration** of stay was long

- *Genetic evidence:*
- Confirms a long history of **endogamy** (group endogamy still current)
- Confirms genetic **distinctness** from coterritorial populations
- Confirms a genetic connection with **India**
- Confirm splits into **dialect groups?**
- Roughly confirm **dates of splits?**

# Is the Roma case special?

- Yes, language as important part of identity
- Yes, one milennium of endogamy
- Yes, they migrated over a long distance
- No, we are all humans