



ΕΛΙΔΕΚ.

Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας



ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ



Πανεπιστήμιο Κρήτης

Μορφο-συντακτικά χαρακτηριστικά των γλωσσών και βιολογικά εργαλεία για τη μελέτη της γλωσσογένεσης



ModelGloss

Πανεπιστήμιο Αιγαίου

Τετάρτη 26 Μαΐου 2021



Αυτή είναι μία εισαγωγή στις σύγχρονες φυλογενετικές μεθόδους στην ιστορική γλωσσολογία και παρουσίαση των πρώτων αποτελεσμάτων του προγράμματος «Μοντελοποιώντας τη Γλωσσογένεση» (ModelGloss) το οποίο χρηματοδοτείται από το ΕΛ.ΙΔ.Ε.Κ.

- Το πρόγραμμα ερευνά τη γλωσσική αλλαγή με μεθόδους εμπνευσμένες από την εξελικτική βιολογία.
- Χρησιμοποιεί ομοιότητες και διαφορές από τη μορφολογία και τη σύνταξη των γλωσσών για να ανιχνεύσει κοινή καταγωγή και γεωγραφικές επαφές μεταξύ τους.
- Συνδυάζει ποιοτικές και ποσοτικές μεθόδους για να αναδείξει τις ιδιαιτερότητες της εξέλιξης των γλωσσών σε σχέση με την εξέλιξη των ειδών.
- Αποτιμά την ιστορική πληροφορία που εμπεριέχουν τα γλωσσικά δεδομένα.



Ερευνητική Ομάδα:

- Έλενα Αναγνωστοπούλου (Πανεπιστήμιο Κρήτης)
- Μανόλης Λαδουκάκης (Πανεπιστήμιο Κρήτης)
- Δημήτρης Μιχελιουδάκης (Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης)
- Μαρία-Μαργαρίτα Μακρή (Πανεπιστήμιο Αιγαίου, Πανεπιστήμιο Κρήτης)
- Πάυλος Παυλίδης (Ίδρυμα Τεχνολογίας και Έρευνας – ΙΤΕ)
- Αθανάσιος-Μιχαήλ Ραμαδανίδης (Πανεπιστήμιο Κρήτης)

Ιστοσελίδα:

<http://modelgloss.philology.uoc.gr/>

Γλωσσολογία και Βιολογία

- Η ιστορική γλωσσολογία από το 19ο αιώνα χρησιμοποιεί για την επανασύνθεση της ιστορικής εξέλιξης των γλωσσών (γλωσσογένεση, glossogeny, Hurford 1990, Fitch 2010) τα εργαλεία που χρησιμοποιεί η εξελικτική βιολογία για την ανακατασκευή της βιολογικής εξέλιξης των ειδών (φυλογένεση, phylogeny).

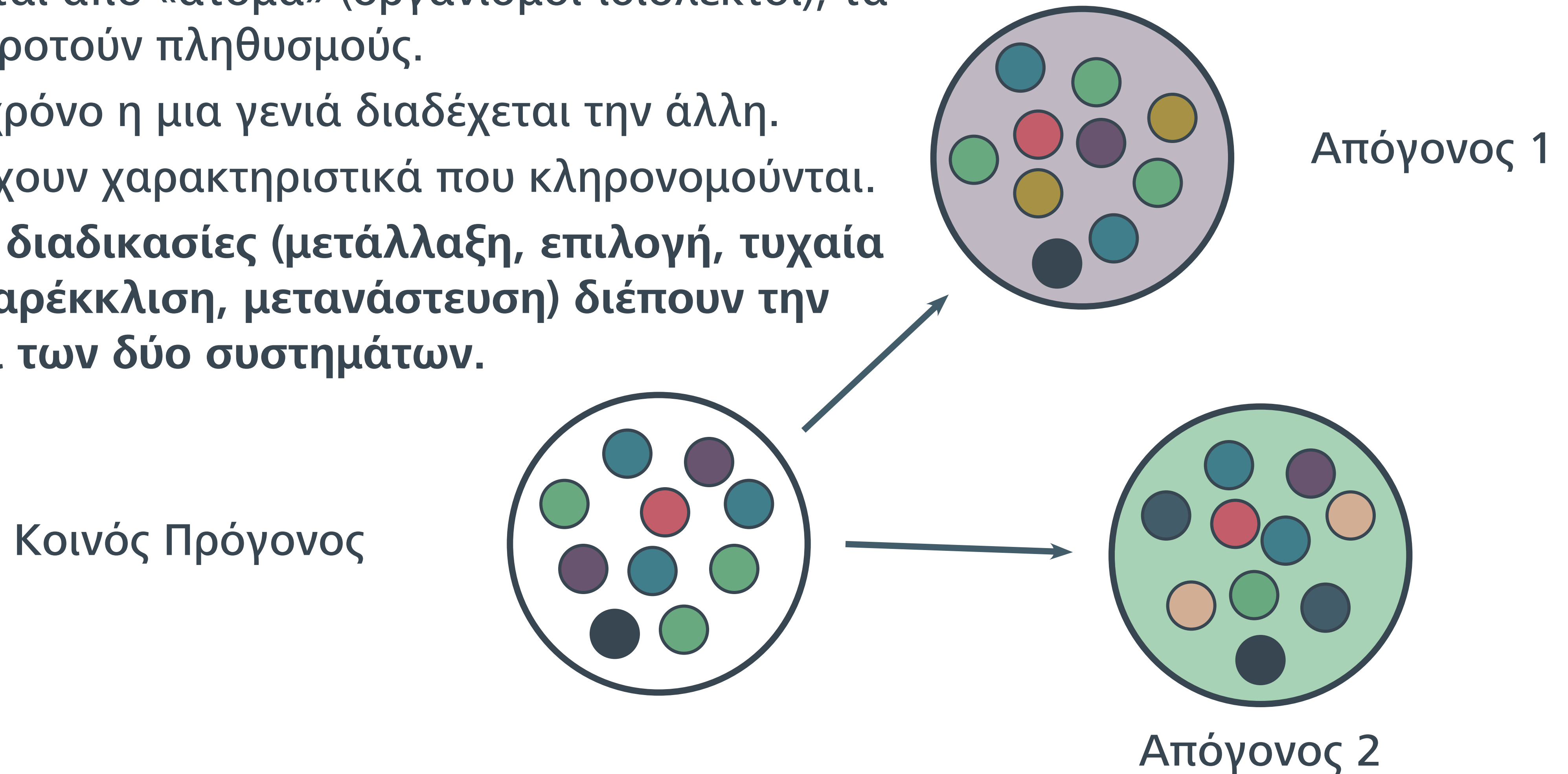
Ιστορική καταγωγή των γλωσσών (γλωσσογένεση)

- Οι γλώσσες, όπως και τα βιολογικά είδη, έχουν κοινούς προγόνους και εξελίχθηκαν στο χρόνο και στο χώρο καθώς οι ανθρώπινοι πληθυσμοί εξαπλώθηκαν στον πλανήτη.
- Όπως και με τα βιολογικά είδη, οι συστηματικές ομοιότητες μεταξύ γλωσσών συχνά αντανακλούν την κοινή καταγωγή τους και οι διαφορές μεταξύ των γλωσσών μεγαλώνουν όσο μεγαλώνει η απόσταση από τους κοινούς προγόνους τους.
- Επομένως, είναι χρήσιμο να μελετηθούν τα δύο εξελίξιμα συστήματα παράλληλα διότι μπορούν να αλληλο-ανατροφοδοτηθούν.

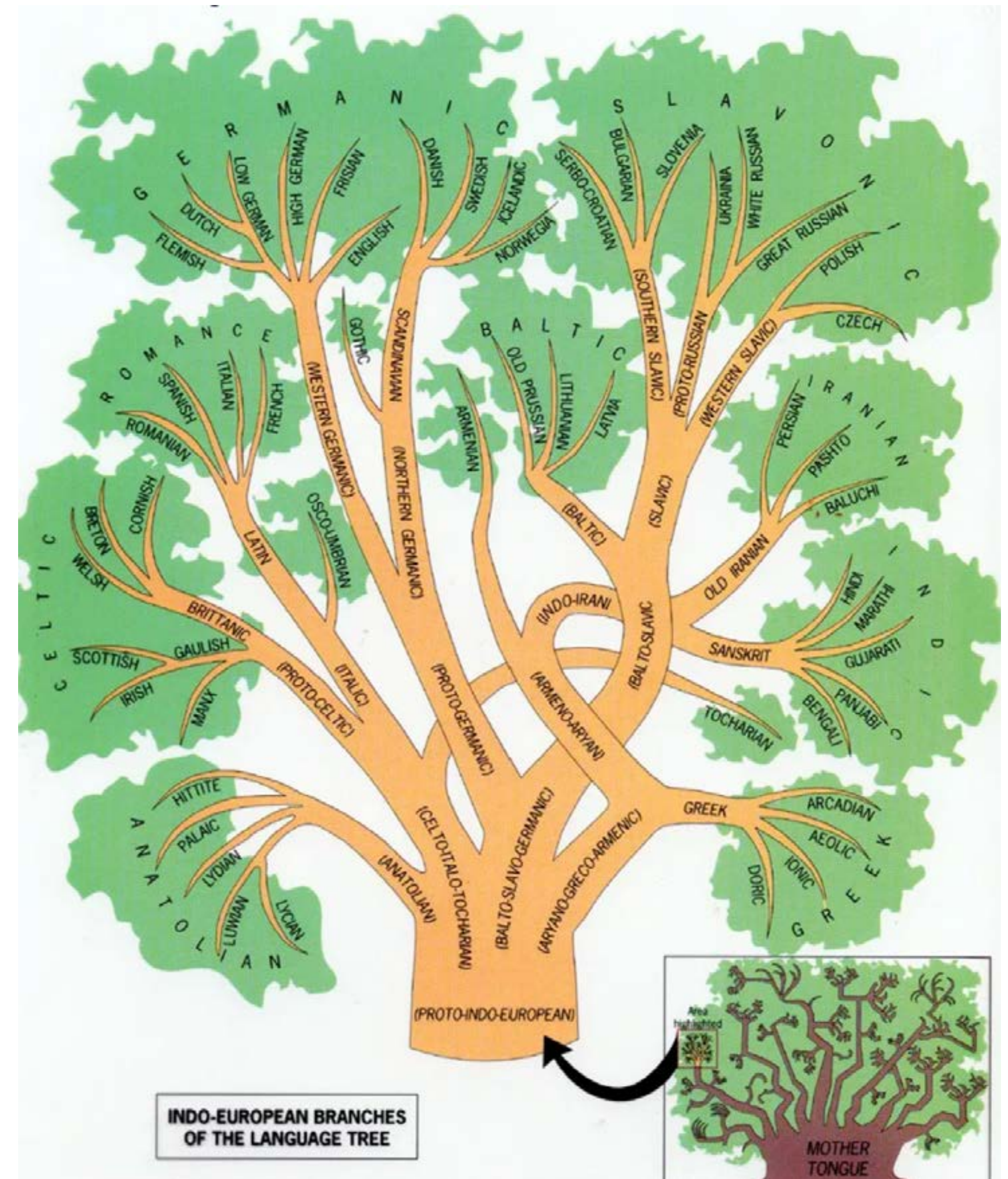
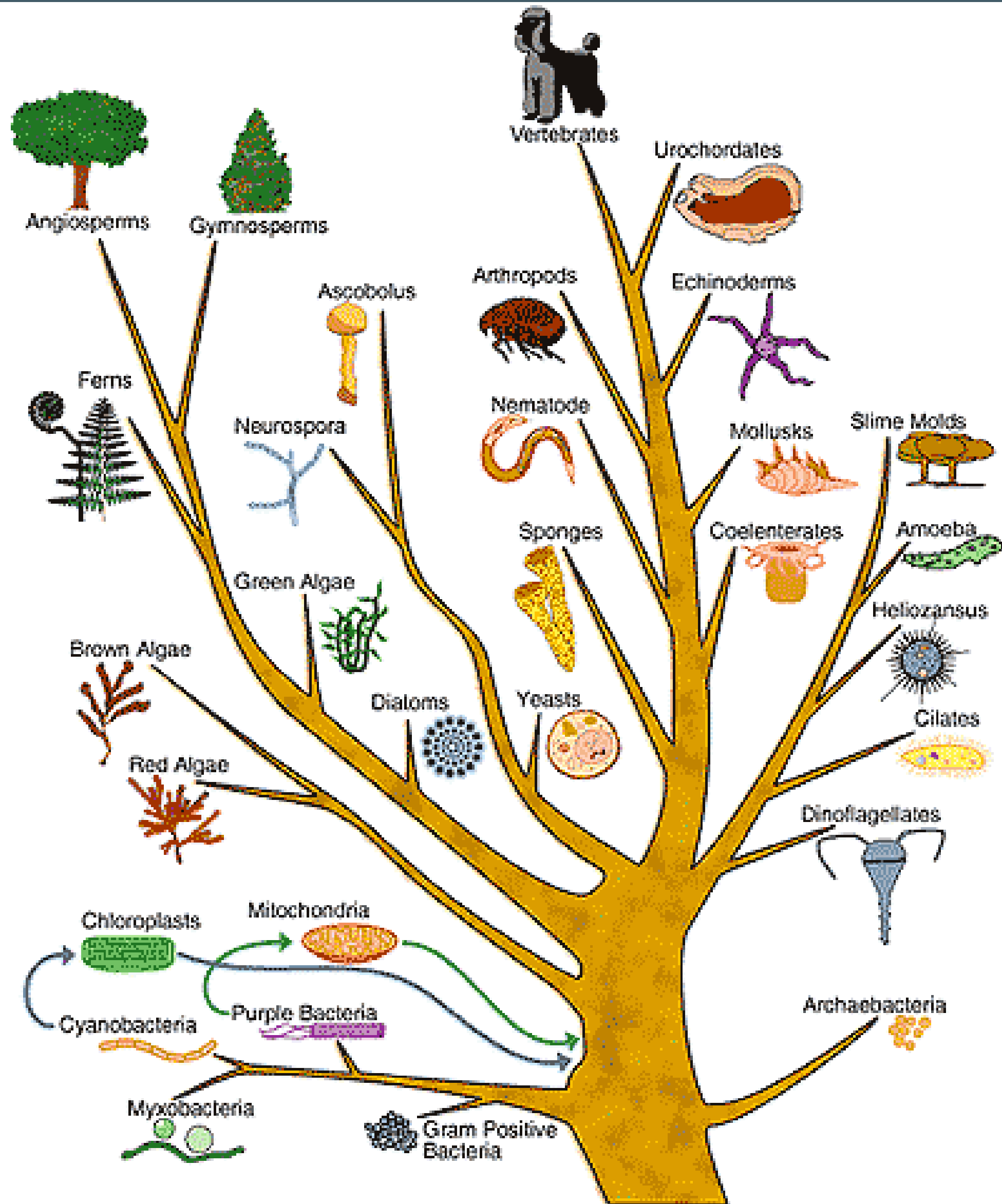
Δύο εξελίξιμα συστήματα:
Οργανισμικά είδη και γλώσσες

Ομοιότητες ανάμεσα στις γλώσσες και τα είδη

- Αποτελούνται από «άτομα» (οργανισμοί-ιδιόλεκτοι), τα οποία συγκροτούν πληθυσμούς.
- Μέσα στο χρόνο η μια γενιά διαδέχεται την άλλη.
- Τα άτομα έχουν χαρακτηριστικά που κληρονομούνται.
- Παρόμοιες διαδικασίες (μετάλλαξη, επιλογή, τυχαία γενετική παρέκκλιση, μετανάστευση) διέπουν την εξέλιξη και των δύο συστημάτων.

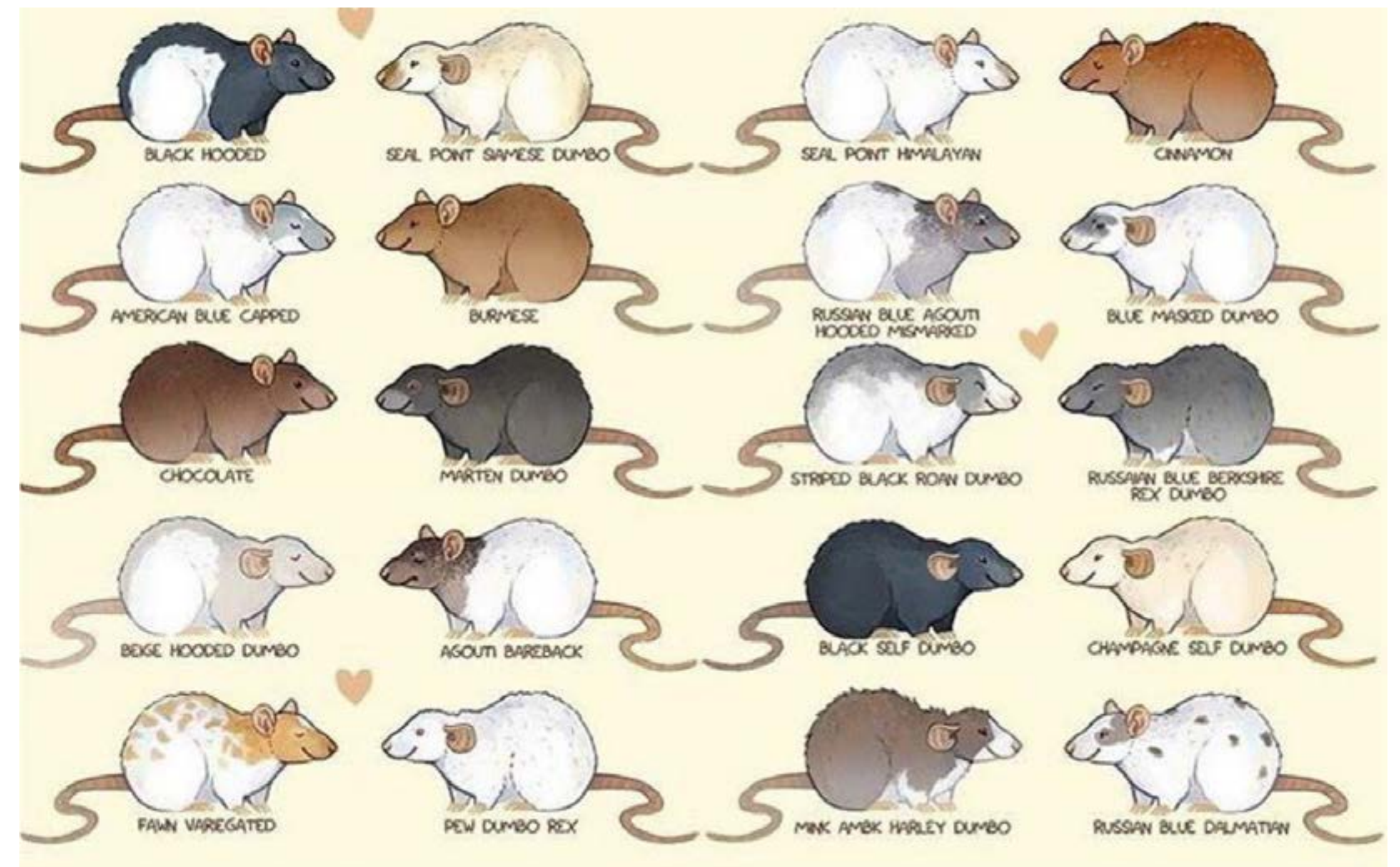
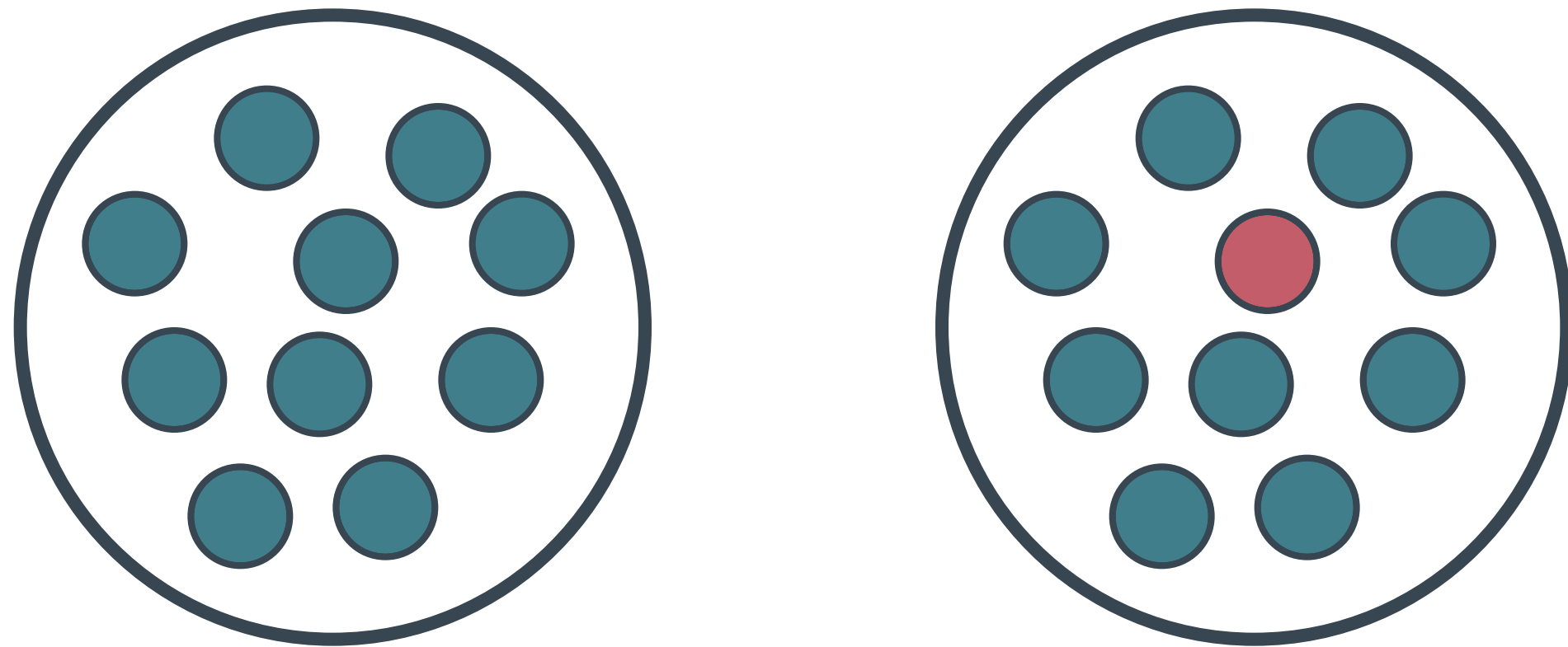


Εξέλιξη είναι η αλλαγή στις κληρονομούμενες ιδιότητες των πληθυσμών



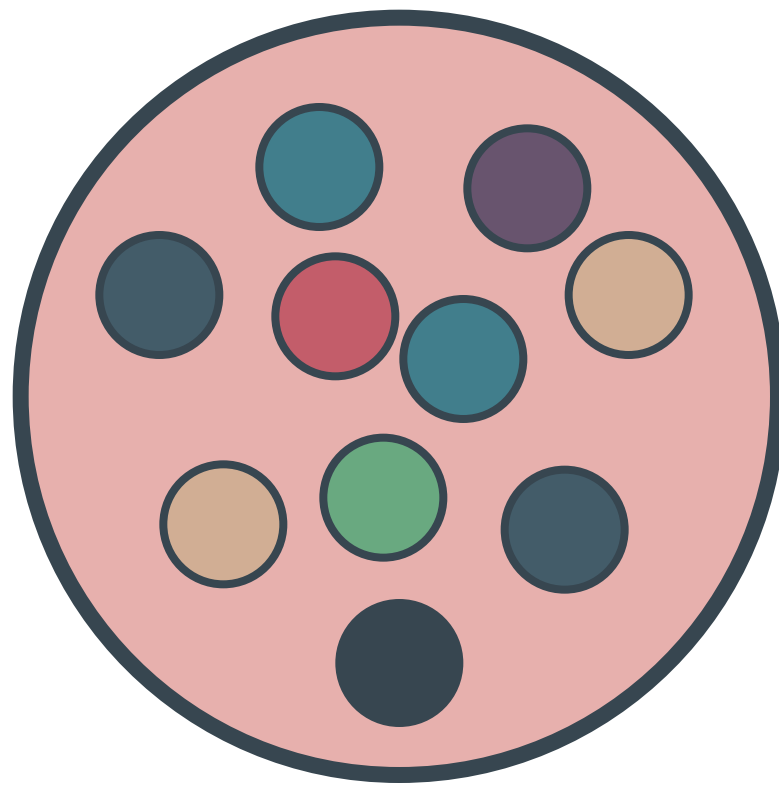
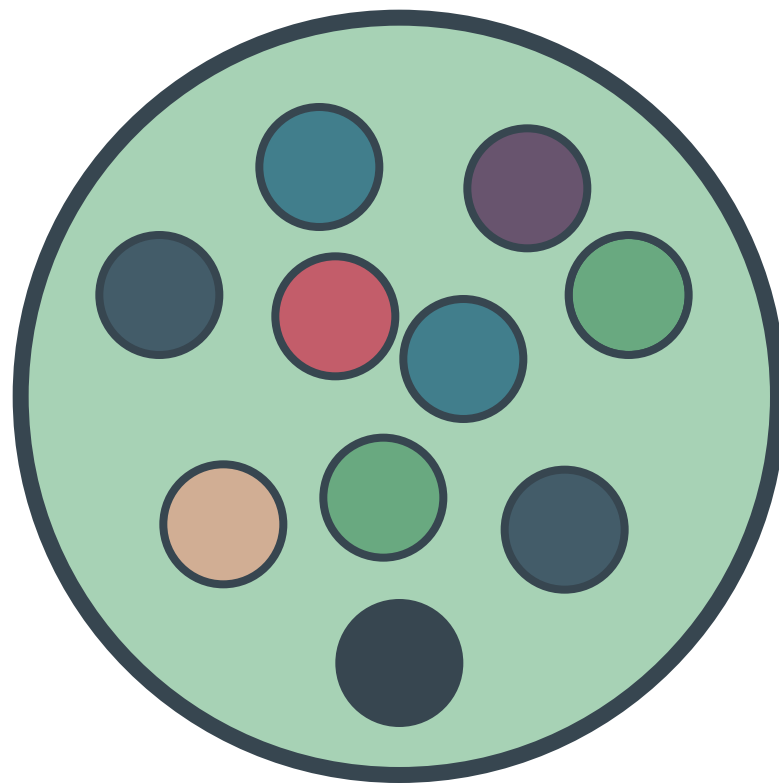
Ποικιλοποίηση της ζωής: Μετάλλαξη

- Η διαδικασία που δημιουργεί de novo ποικιλομορφία.



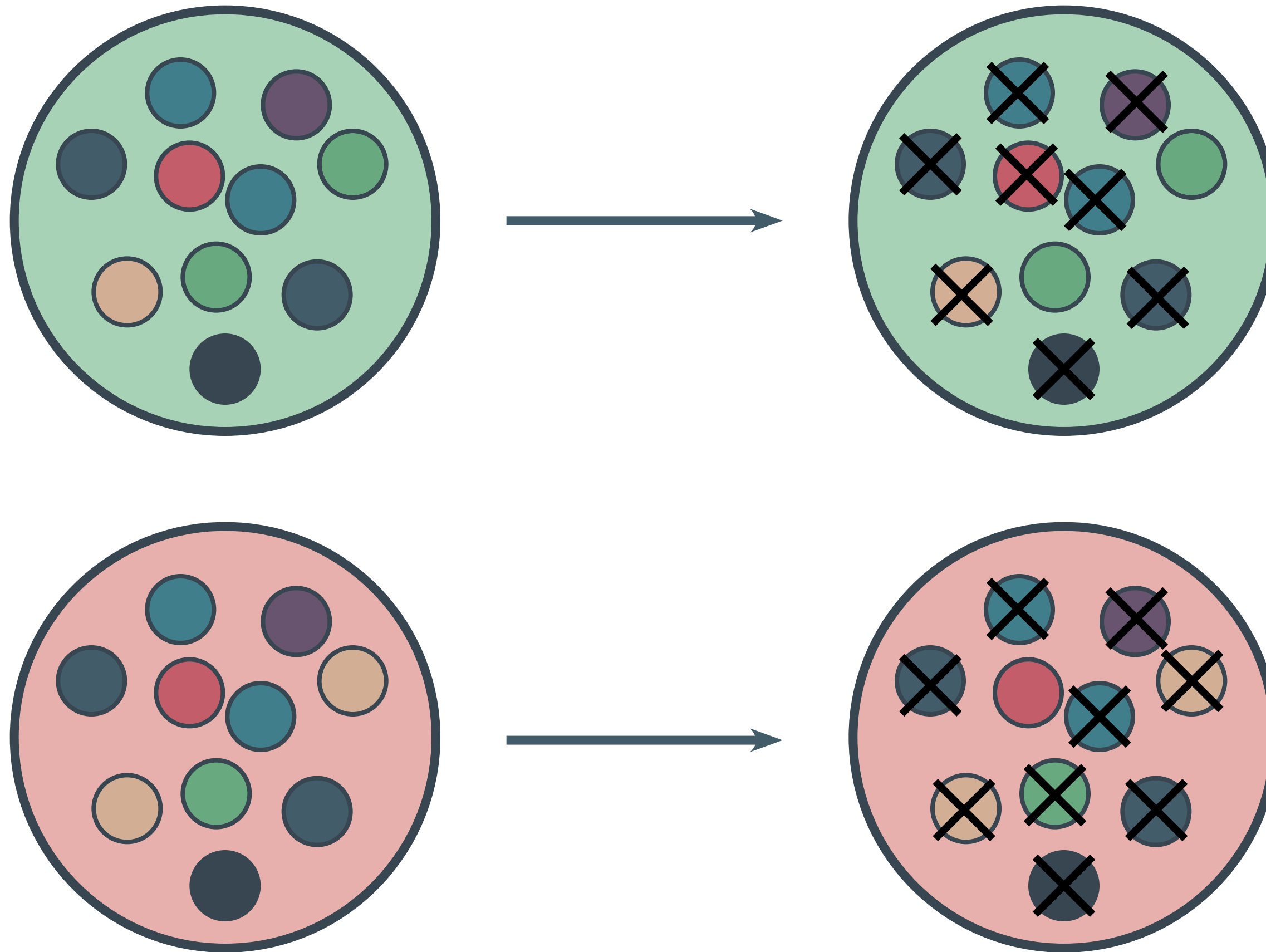
Ποικιλοποίηση της ζωής: Φυσική επιλογή

- Μη τυχαία μεταβολή



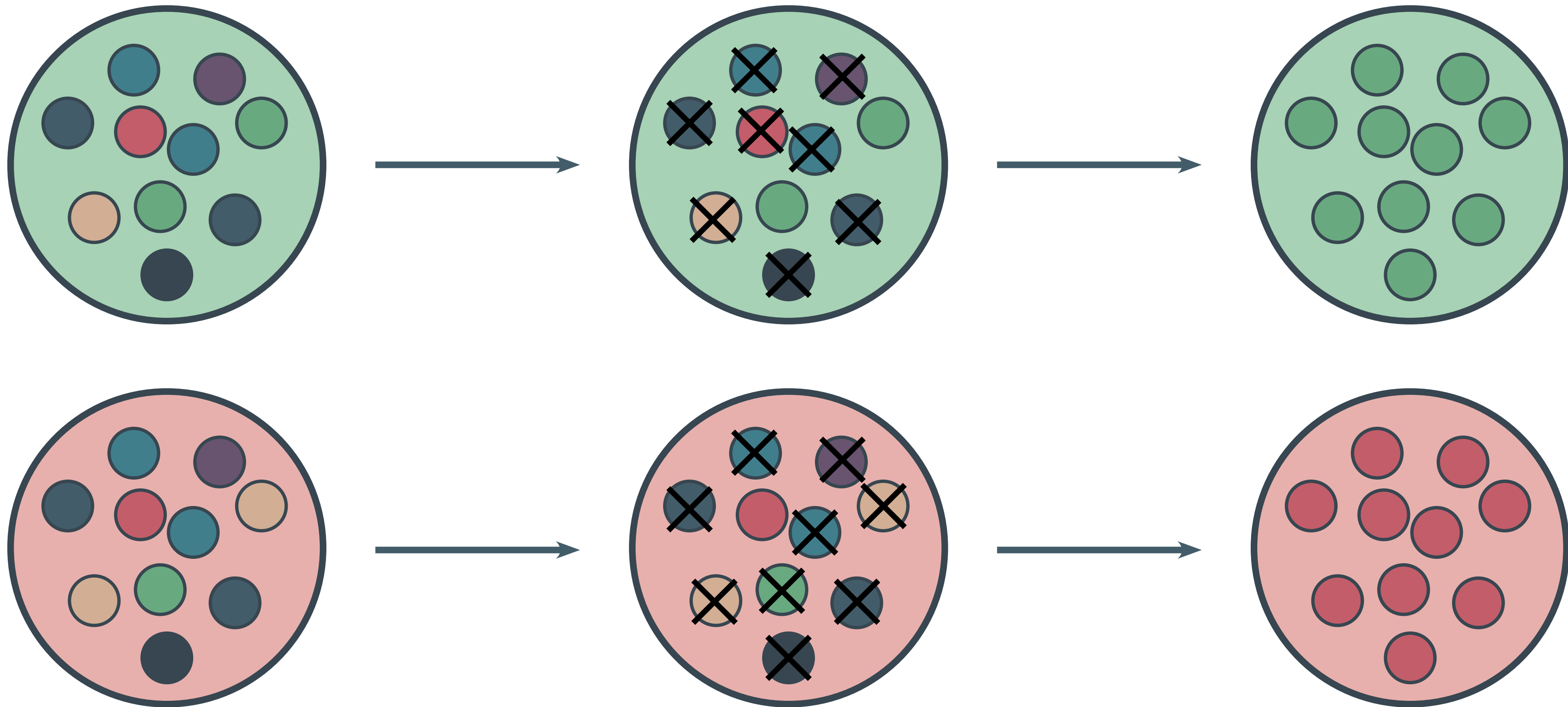
Ποικιλοποίηση της ζωής: Φυσική επιλογή

- Μη τυχαία μεταβολή



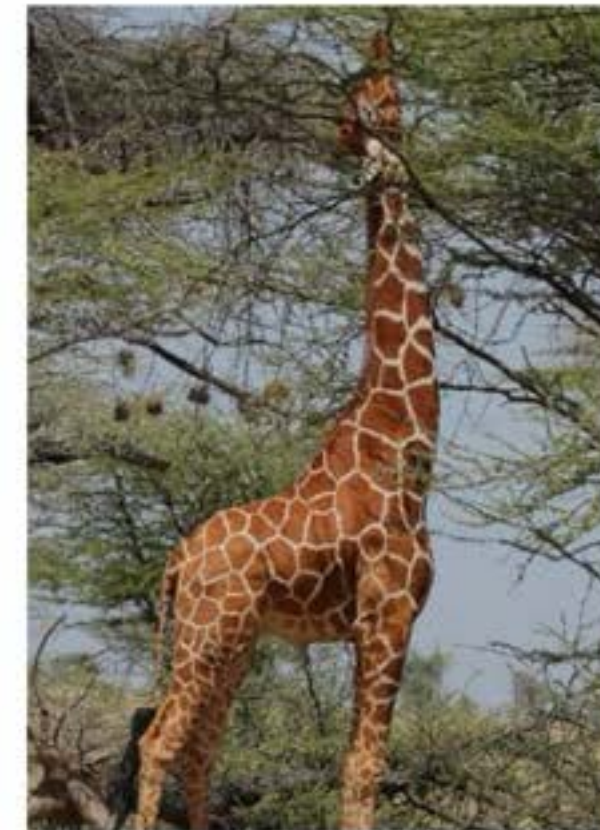
Ποικιλοποίηση της ζωής: Φυσική επιλογή

- Μη τυχαία μεταβολή

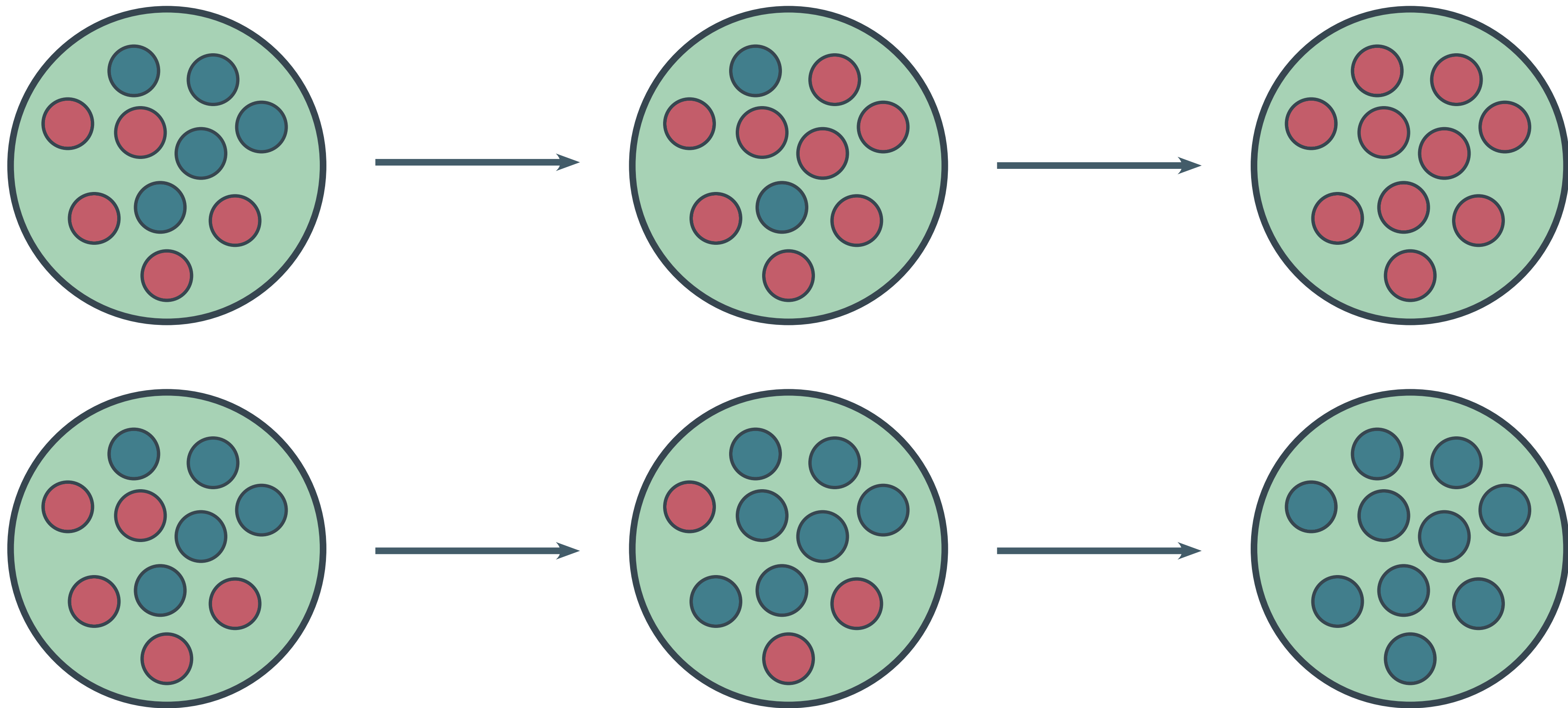


Ποικιλοποίηση της ζωής: Φυσική επιλογή

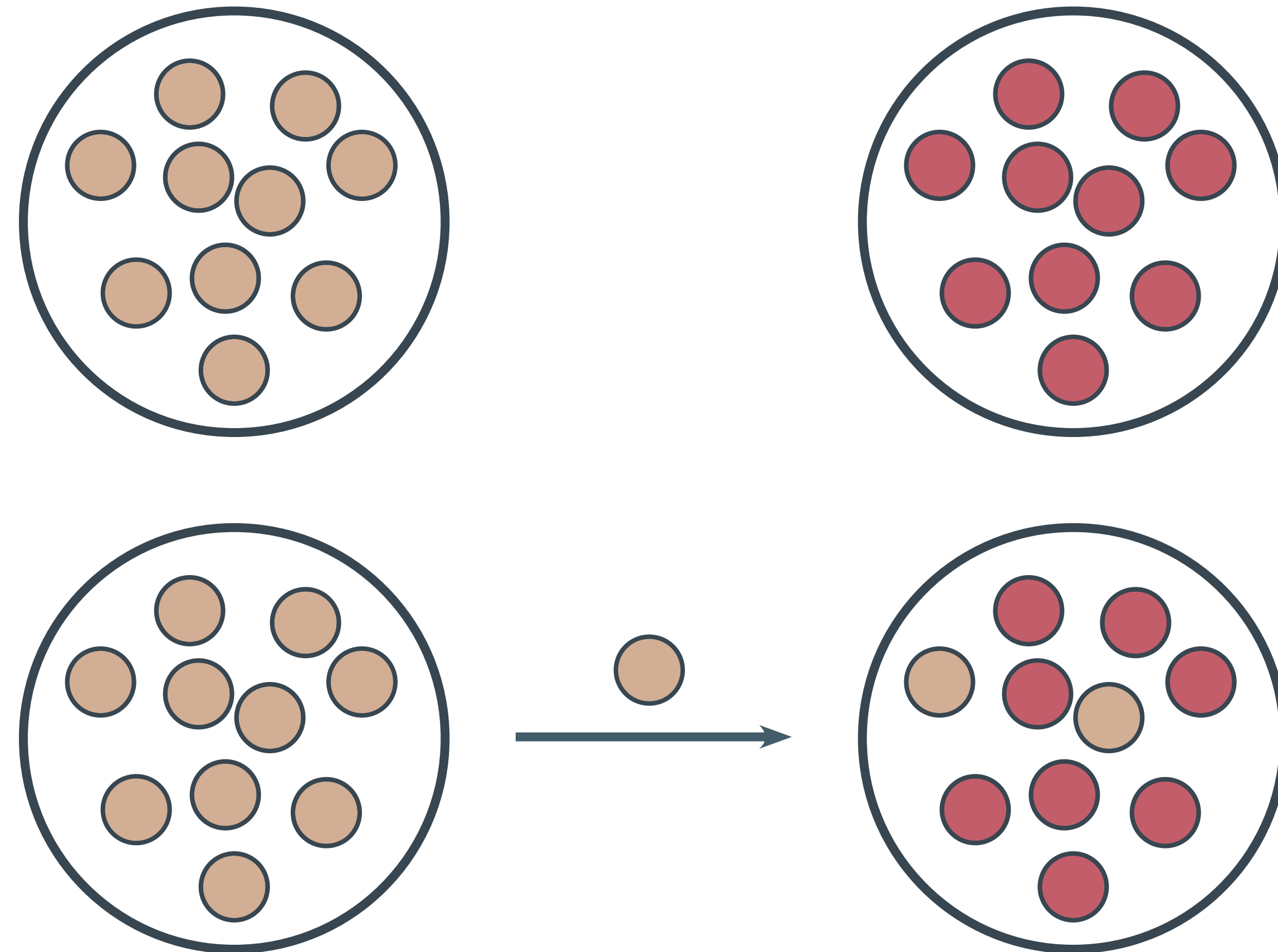
Προσαρμογές



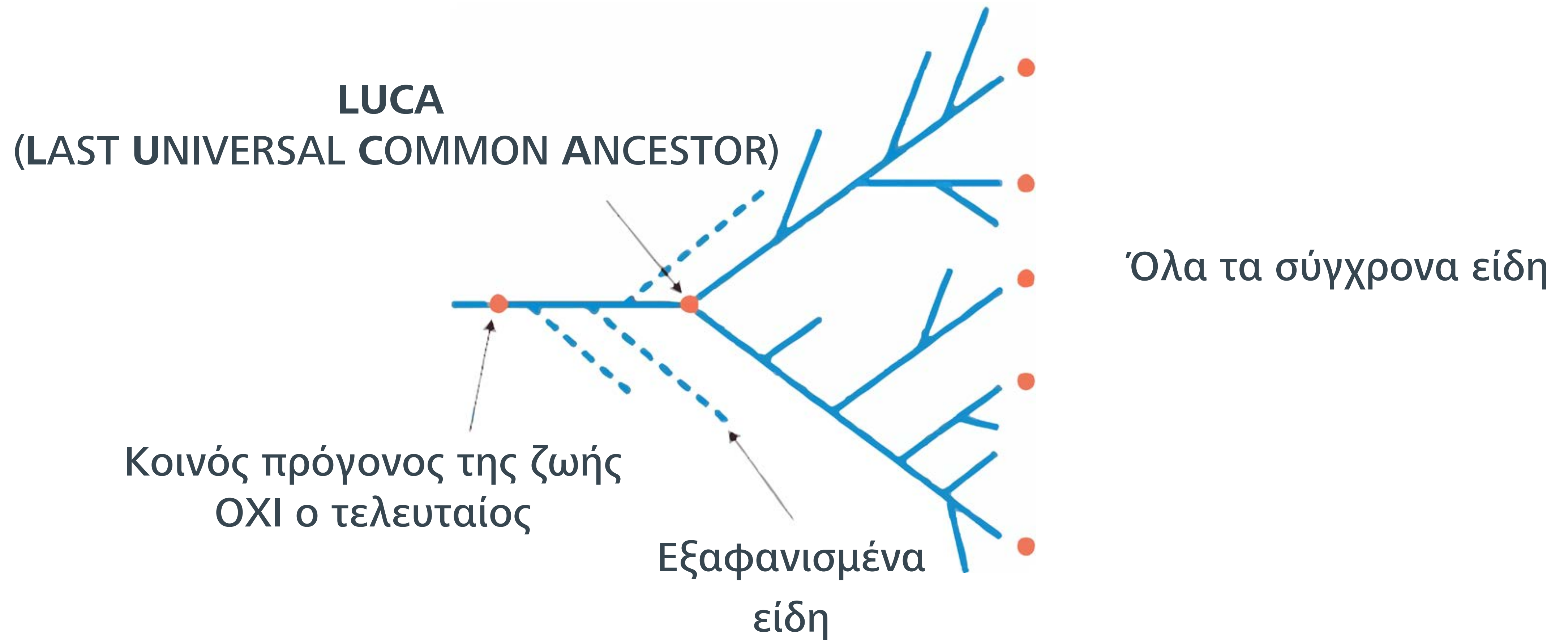
Ποικιλοποίηση της ζωής: Τυχαία γενετική παρέκκλιση



Ποικιλοποίηση της ζωής: Μετανάστευση



Η ζωή έχει κοινή καταγωγή

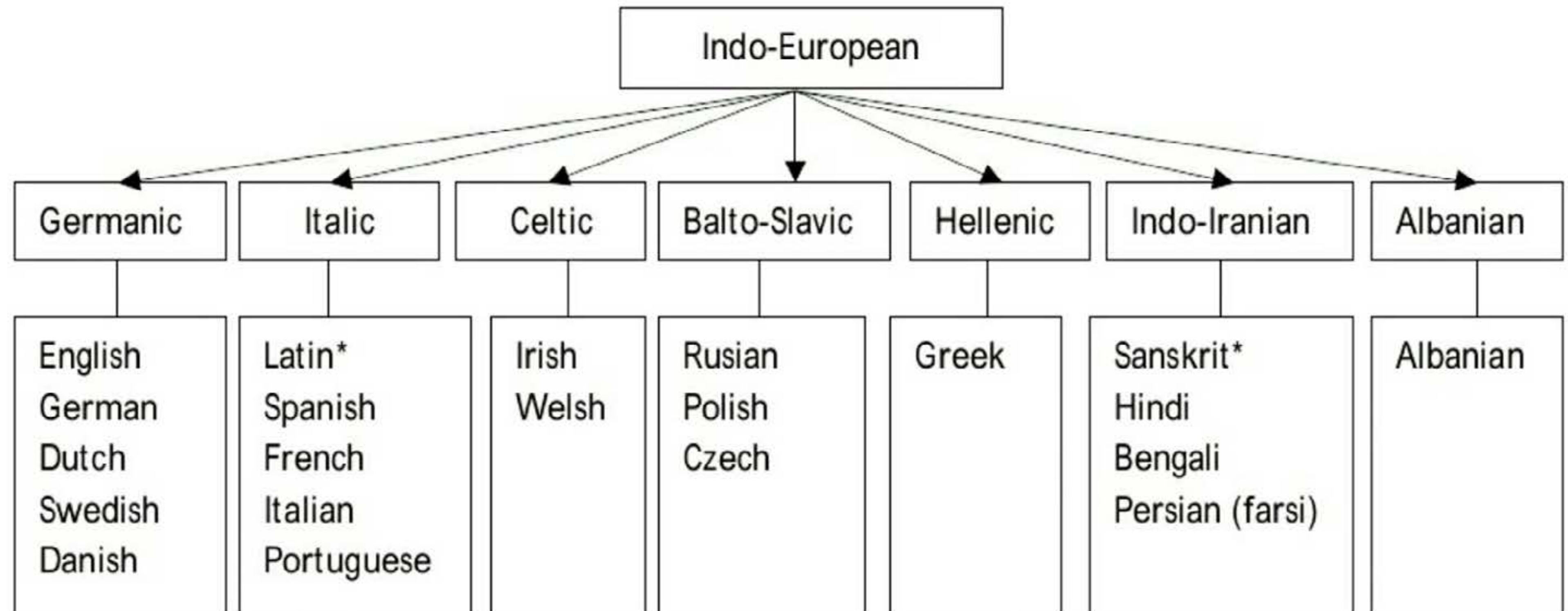


Οι γλώσσες εξελίσσονται με αντίστοιχες διαδικασίες

Οικογένεια

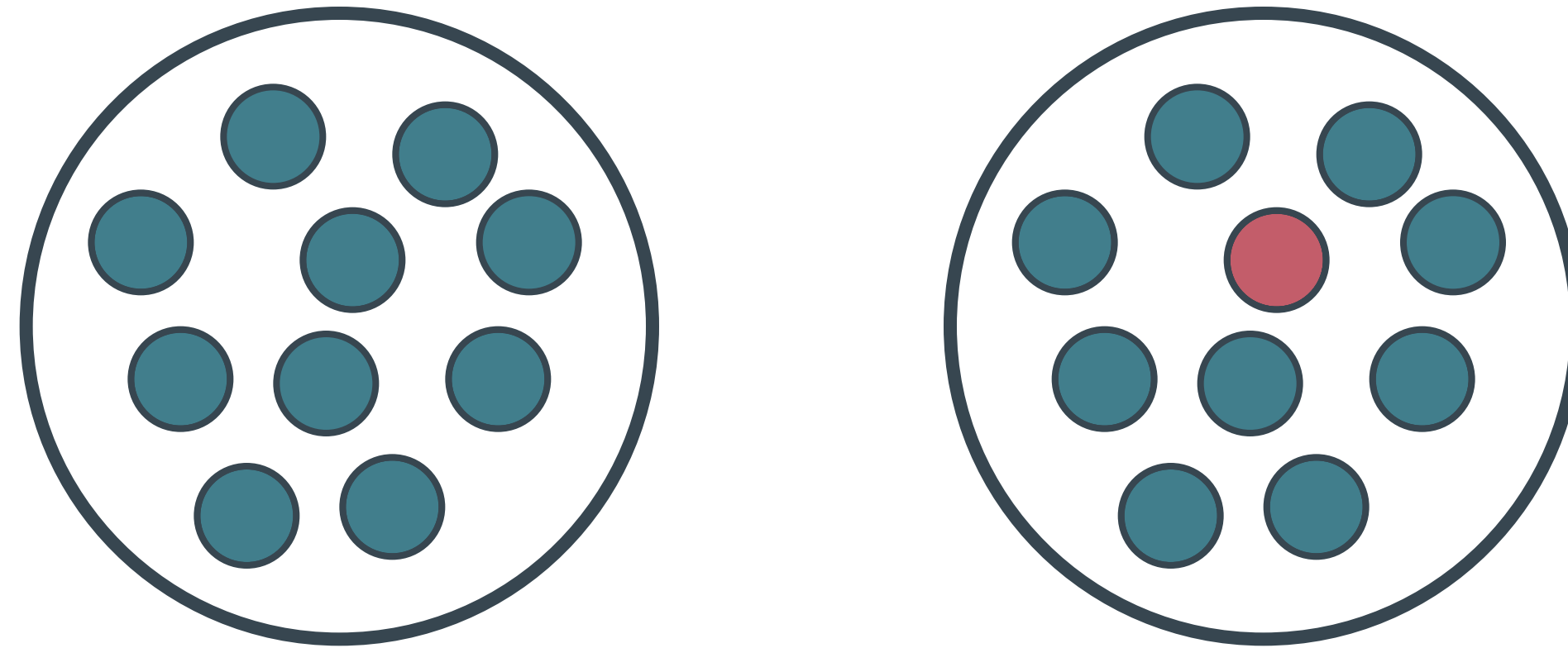
Γένη/Κλάδοι

Γλώσσες



Εξέλιξη γλωσσών: Δημιουργία ποικιλομορφίας

- Καινοτομία (Innovation)
- Actuation problem



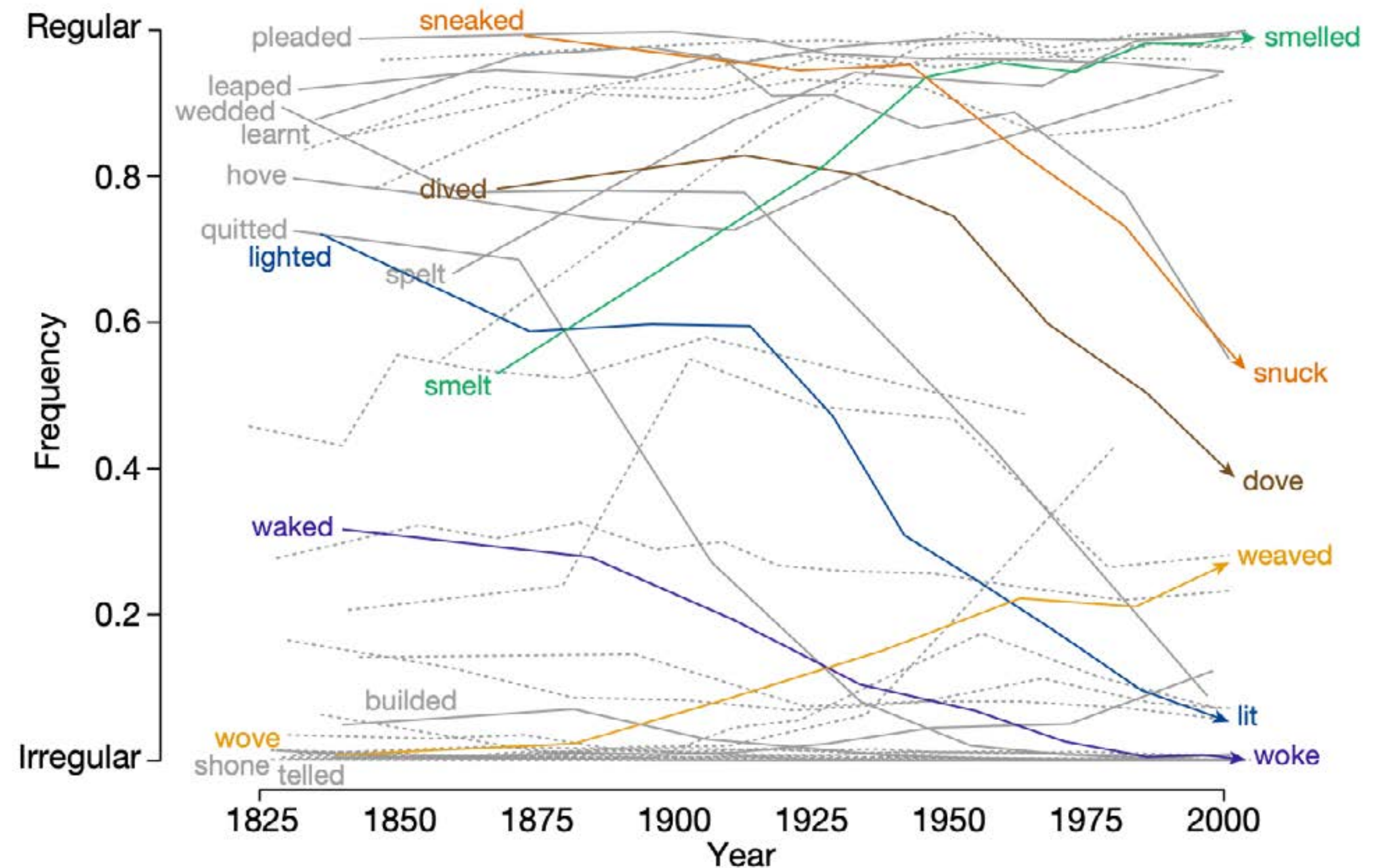
Εξέλιξη γλωσσών: Επιλογή ή τυχαιότητα

- drive – drove
- dive – dived – dove
- επιλογή του dove στο πρότυπο του drove που είχε μεγάλη συχνότητα στις αρχές του 20ου αι.

VS

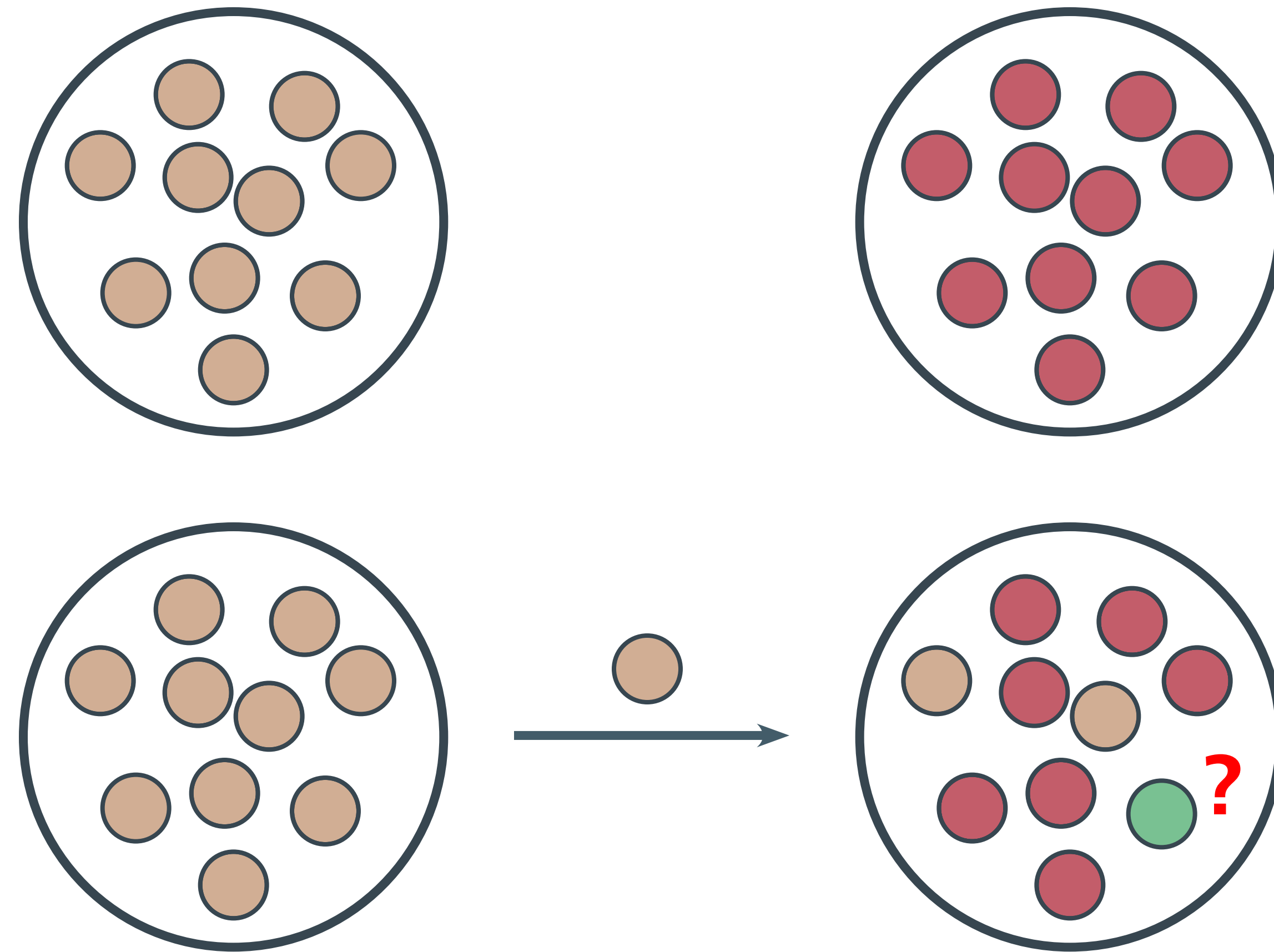
- quitted – quit
- learned – learnt

Ανωμαλοποίηση ή ομαλοποίηση ρημάτων



Newberry et al 2017 Nature

Εξέλιξη γλωσσών: Γλωσσική επαφή



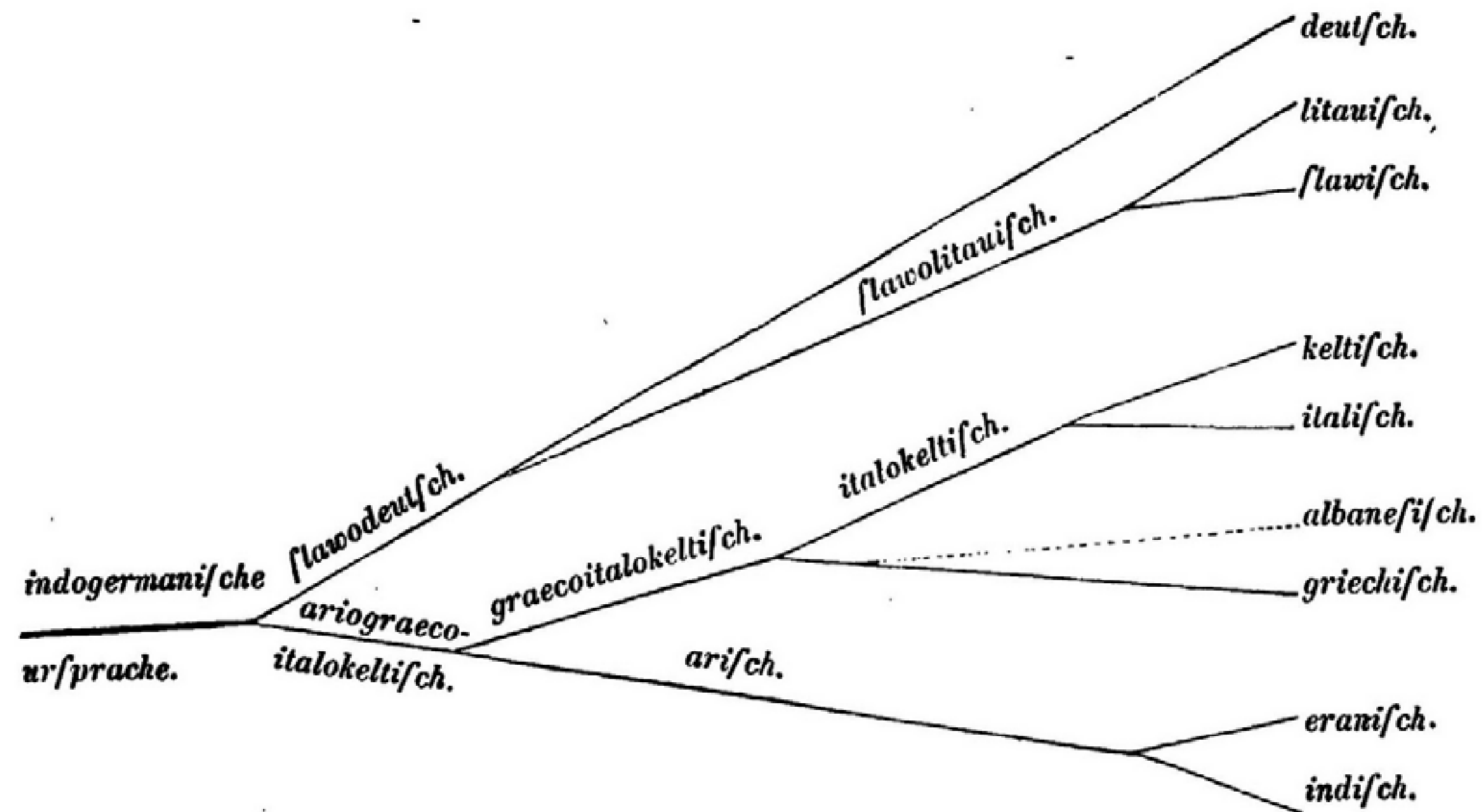
Διαφορές (μερικές) των δύο συστημάτων

Έννοια	Βιολογία	Γλώσσα
Κληρονόμηση	Κάθετη	Κάθετη και οριζόντια
Γονείς	1–2	1 – πολλοί
Κληρονόμηση επίκτητων χαρακτηριστικών	Όχι	Ναι
Ανταλλαγή πληροφορίας	Μεταξύ ειδών όχι	Μεταξύ γλωσσών ναι

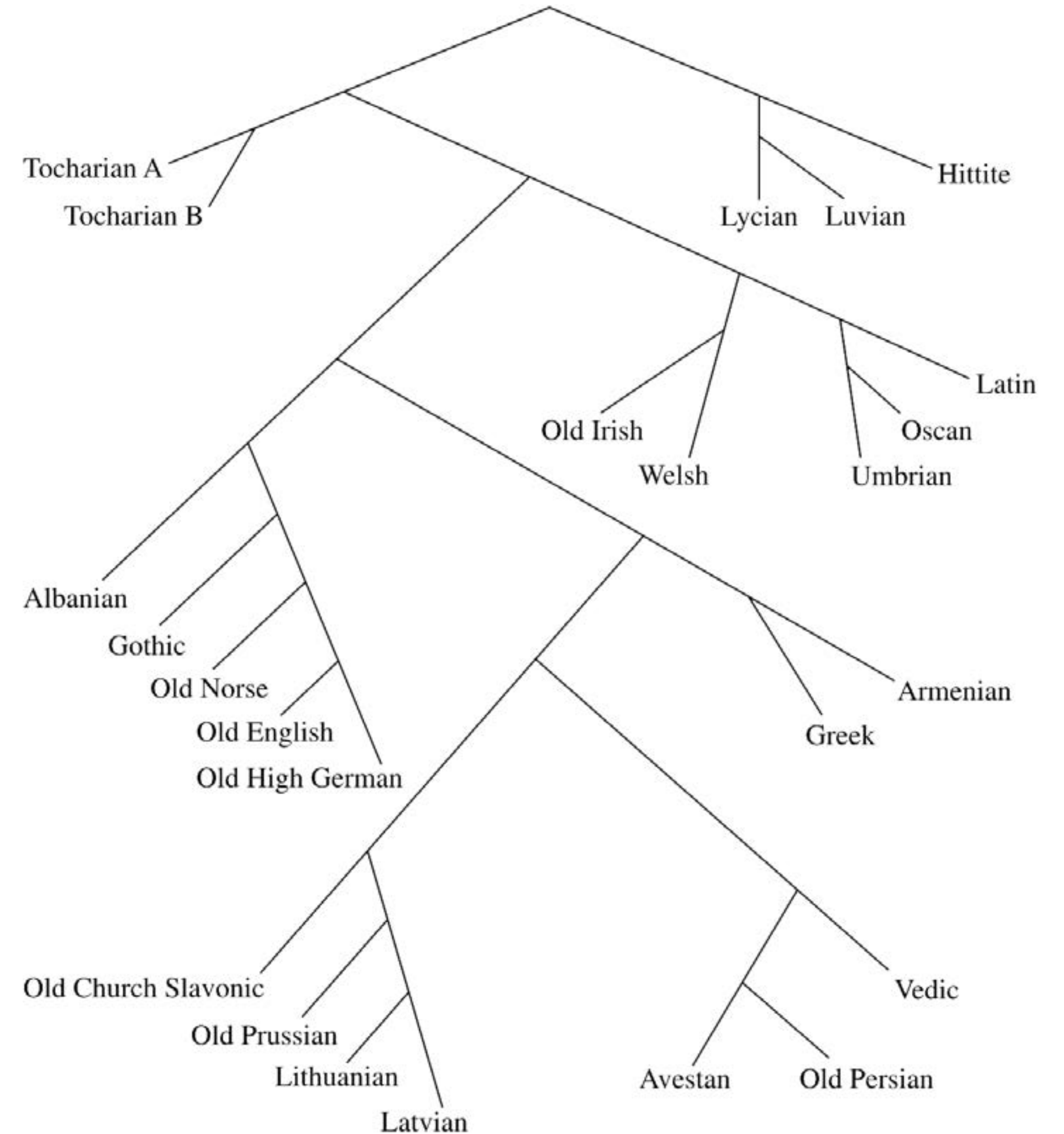
Η γλώσσα ως εξελίξιμο σύστημα: Ερωτήματα

- Πώς συγκρίνουμε και ταξινομούμε τις γλώσσες;
 - Ποια δεδομένα και ποιες μεθόδους χρησιμοποιούμε;
- Πολύ συνηθισμένη μέθοδος, δανεισμένη από τη Βιολογία, είναι τα δέντρα

Παραδοσιακή και υπολογιστική φυλογενετική

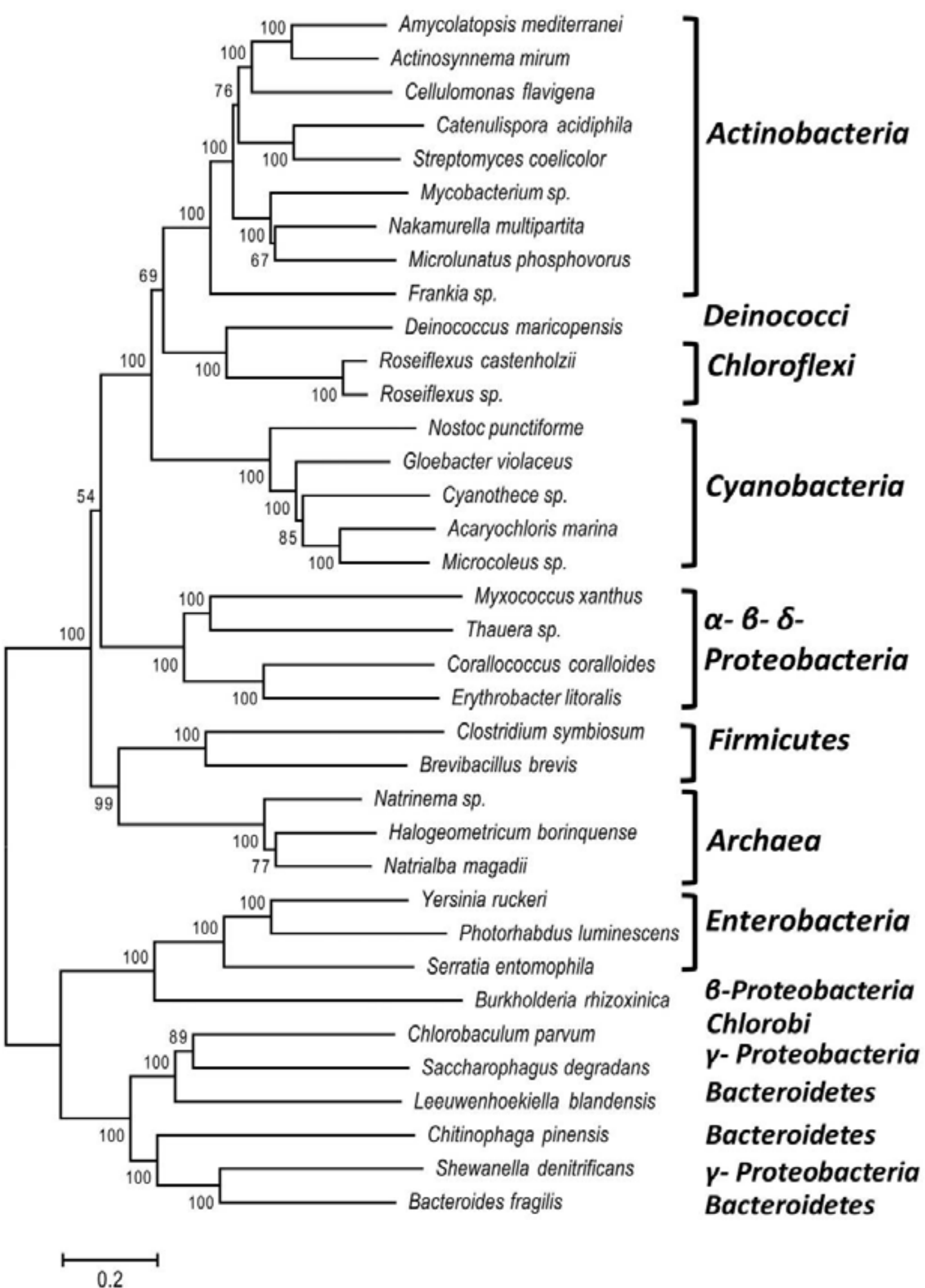
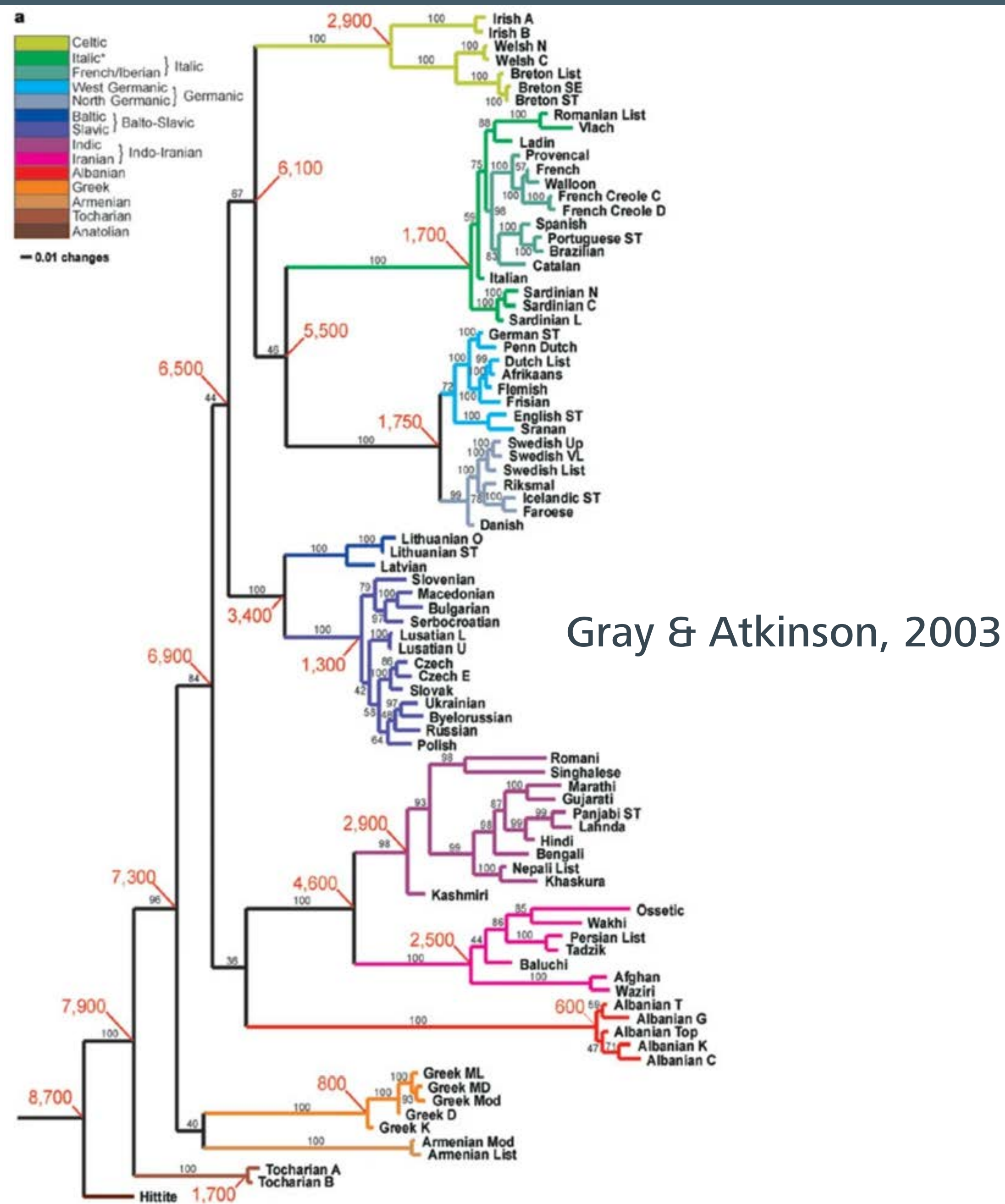


Schleicher, 1863



Ringe, Warnow, & Taylor, 2002

Εκτίμηση γενετικών σχέσεων με φυλογενετικές μεθόδους



Sarris, Ladoukakis et al., 2014

Η σημασία της επιλογής των
κατάλληλων χαρακτηριστικών

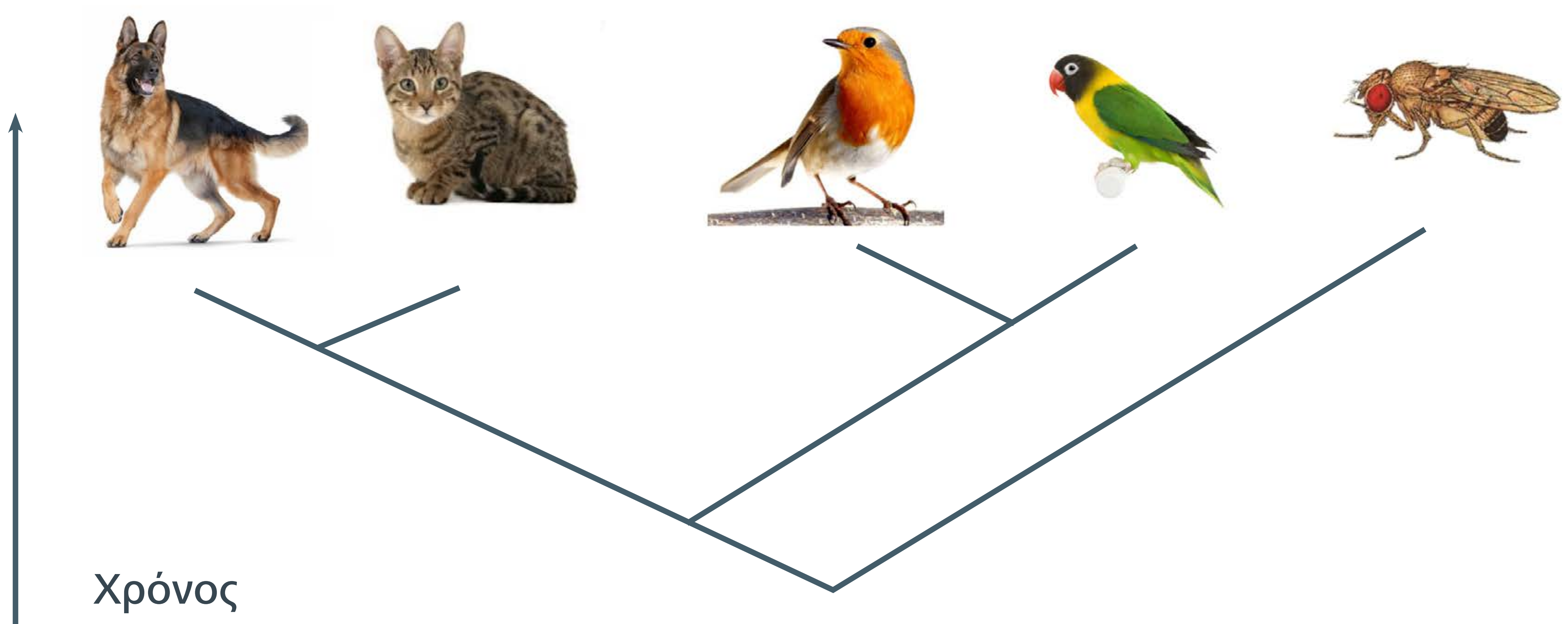
Σχέσεις μεταξύ των οργανισμών



Σχέσεις μεταξύ των οργανισμών



Σχέσεις μεταξύ των οργανισμών



Χαρακτήρες και καταστάσεις χαρακτήρων

	Συμμετρία	Αριθμός άκρων	Μεταμερή	Οστά	Κάλυψη σώματος	Πρόσληψη τροφής	Αυγά	Φτερά
Σκύλος	Αμφίπλευρη	4	Όχι	Ναι	Τρίχες	Στόμα	Όχι	Όχι
Γάτα	Αμφίπλευρη	4	Όχι	Ναι	Τρίχες	Στόμα	Όχι	Όχι
Κοκκινολαίμης	Αμφίπλευρη	4	Όχι	Ναι	Πτέρωμα	Ράμφος	Ναι	Ναι
Παπαγάλος	Αμφίπλευρη	4	Όχι	Ναι	Πτέρωμα	Ράμφος	Ναι	Ναι
Μύγα	Αμφίπλευρη	6	Ναι	Όχι	Σμύριγγες	Μυζητήρας	Ναι	Ναι

Χαρακτήρες και καταστάσεις χαρακτήρων

	Συμμετρία	Αριθμός άκρων	Μεταμερή	Οστά	Κάλυψη σώματος	Πρόσληψη τροφής	Αυγά	Φτερά
Σκύλος	Αμφίπλευρη	4	Όχι	Ναι	Τρίχες	Στόμα	Όχι	Όχι
Γάτα	Αμφίπλευρη	4	Όχι	Ναι	Τρίχες	Στόμα	Όχι	Όχι
Κοκκινολαίμης	Αμφίπλευρη	4	Όχι	Ναι	Πτέρωμα	Ράμφος	Ναι	Ναι
Παπαγάλος	Αμφίπλευρη	4	Όχι	Ναι	Πτέρωμα	Ράμφος	Ναι	Ναι
Μύγα	Αμφίπλευρη	6	Ναι	Όχι	Σμύριγγες	Μυζητήρας	Ναι	Ναι

Σχέσεις μεταξύ των οργανισμών



Σχέσεις μεταξύ των οργανισμών



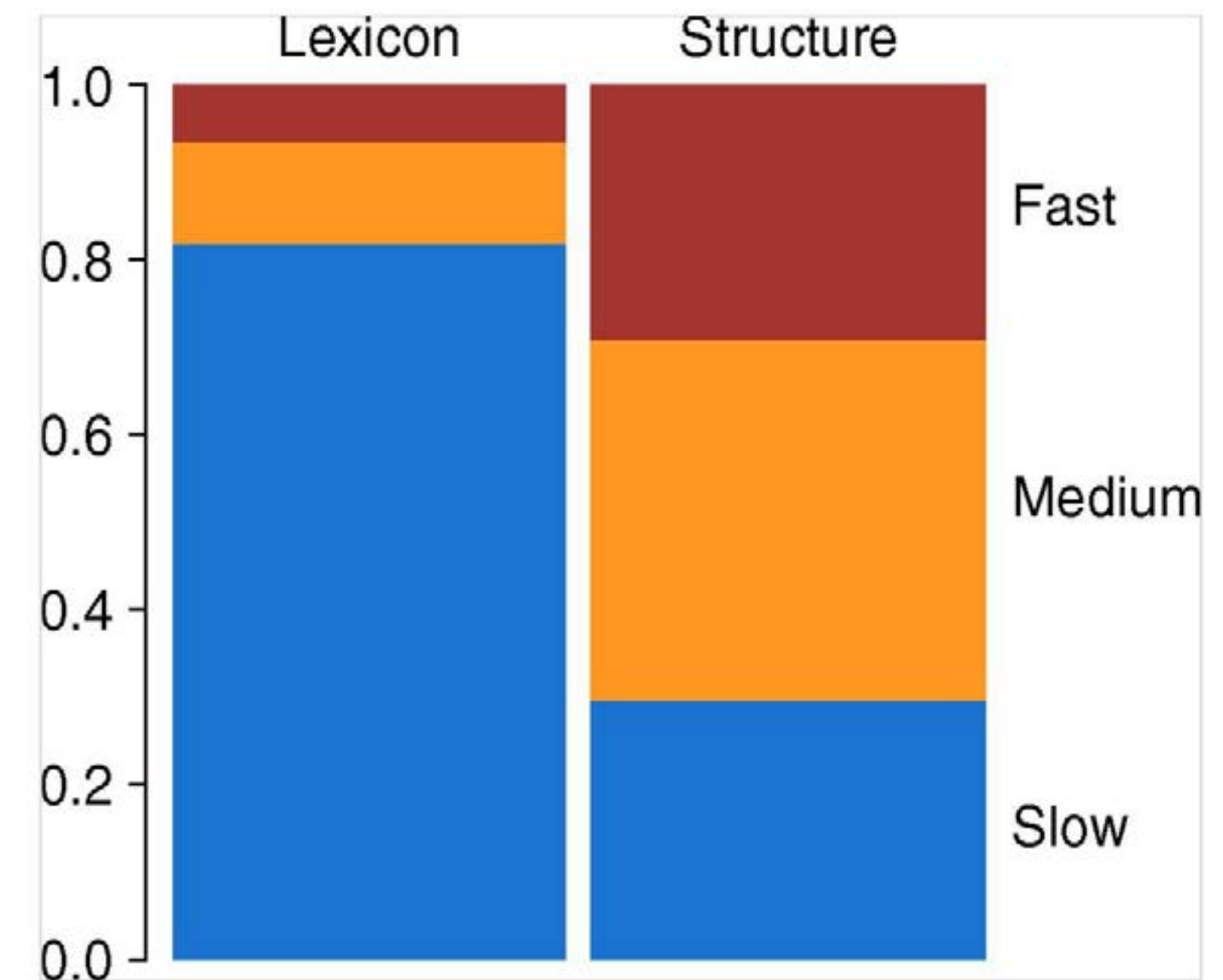
Ομόρριζα vs μορφοσυντακτικά δεδομένα I

- Στη σύγχρονη ιστορική έρευνα που στηρίζεται σε φυλογενετικές μεθόδους, τα δεδομένα που χρησιμοποιούνται προέρχονται από το συντηρητικό λεξιλόγιο των γλωσσών, το οποίο έχει αποδειχθεί ότι αντιστέκεται στο δανεισμό (Swadesh 1955).
- Σύμφωνα με την κρατούσα άποψη στην ιστορική γλωσσολογία, αυτή η μέθοδος μπορεί να φτάσει σε ένα ιστορικό βάθος 6000–10000 χρόνων (πβλ. όμως Gray & Atkinson 2003; Jäger 2015).
- Η Nichols (1992) προτείνει ότι τα γραμματικά χαρακτηριστικά μπορούν να μας βοηθήσουν να ανιχνεύσουμε βαθύτερες ιστορικές σχέσεις, τουλάχιστον 15.000 ετών ενδεχομένως έως και 50.000 ετών.

Αγγλικά	Ιταλικά	Ρωσικά	Ουγγρικά
one	uno	odin	egy
two	due	dva	két
three	tre	tri	három
mother	madre	mat'	anya
brother	fratello	brat	testvér
sister	sorella	sestrá	nővér

Ομόρριζα vs μορφοσυντακτικά δεδομένα II

- Από την άλλη μεριά, οι Greenhill et al. (2017) στηρίζουν ότι τα γραμματικά χαρακτηριστικά είναι λιγότερο αξιόπιστα για δύο πιθανούς λόγους:
 - i. πολλά γραμματικά χαρακτηριστικά αλλάζουν γρηγορότερα από το συντηρητικό τμήμα του λεξικού,
 - ii. τα δομικά δεδομένα εμφανίζουν υψηλότερα επίπεδα αντιφατικού σήματος λόγω
 - › παράλληλης εξέλιξης και
 - › γεωγραφικής διάχυσης.



Greenhill et al 2017, PNAS

Ομόρριζα vs μορφοσυντακτικά δεδομένα III

- Οι Greenhill et al. παρατηρούν ότι τα χαρακτηριστικά που εξελίσσονται με μικρότερο ρυθμό φαίνεται να είναι πιο αφηρημένα και λιγότερο διαθέσιμα στην αντίληψη των ομιλητών.
- Στο ίδιο πνεύμα εν μέρει, οι Ringe, Warnow and Taylor (2002) ενσωματώνουν γραμματικούς χαρακτήρες για να βελτιώσουν την φυλογενετική επανασύνθεση της Ινδο-Ευρωπαϊκής.
- Οι Longobardi και Guardiano (2009), Longobardi et al. (2016) χτίζουν φυλογενετικά δέντρα βασιζόμενοι αποκλειστικά σε μορφοσυντακτικά δεδομένα των ονοματικών φράσεων τα οποία παρουσιάζονται ως ένα δίκτυο αλληλεξαρτώμενων παραμέτρων.

Η προσέγγιση του ModelGloss

- Το ModelGloss διερευνά τη δυναμική και τις προκλήσεις των μορφοσυντακτικών δεδομένων.
- Στόχος είναι να διερευνήσουμε και να αποκωδικοποιήσουμε το σήμα στη μορφοσύνταξη, διακρίνοντας:
 1. χαρακτηριστικά που διατηρούν το ιστορικό σήμα από
 2. χαρακτηριστικά που είναι ευαίσθητα στη μεταφορά λόγω γλωσσικής επαφής, και
 3. ομοπλαστικά χαρακτηριστικά.
- Η κύρια υπόθεσή μας είναι ότι:
 - › Τα χαρακτηριστικά στην κατηγορία 1 σχετίζονται με μακρο-/μέσο-παραμέτρους,
 - › Τα χαρακτηριστικά στην κατηγορία 2 σχετίζονται με μικρο-/νάνο-παραμέτρους (βλ. Biberauer et al. 2014; Roberts 2019),
 - › Τα χαρακτηριστικά στην κατηγορία 3 αποκαλύπτουν καθολικές τάσεις.
- Στόχος της ανάλυσης είναι να δώσει καινούργια εργαλεία για τη διάγνωση ιεραρχικών εξαρτήσεων μεταξύ χαρακτήρων.

Η προσέγγιση του ModelGloss

Στην έρευνά μας μελετάμε:

- A) Την ανακατασκευή της ιστορίας των γλωσσών με βάση χαρακτήρες από όλο το φάσμα της μορφο-σύνταξης.
- B) Ελέγχουμε τη **σχέση/εξάρτηση** μεταξύ χαρακτήρων.
 - Χρησιμοποιούμε **στατιστικές (αμερόληπτες) μεθόδους** για να βρούμε την αλληλεξάρτηση χαρακτήρων.
 - Ελέγχουμε τον τρόπο που οι εξαρτήσεις χαρακτήρων επηρεάζουν τα δέντρα (διαφορετικά δέντρα ανάλογα με τον **βαθμό εξάρτησης** των χαρακτήρων)
 - Μέσα από τον στατιστικό έλεγχο της αλληλεξάρτησης των χαρακτήρων, αποκαλύπτουμε **σχέσεις εξάρτησης** που δεν ήταν προφανείς πριν και θα μας οδηγήσουν σε **ιεραρχίες χαρακτήρων**.

Ιεραρχίες χαρακτήρων

Μακροπαράμετροι



Νανοπαράμετροι

Γραμματικός αριθμός;

OXI

NAI

Γένος;

Μαρκάρισμα πεπερασμένων όρων;

OXI

NAI

Γένος βασισμένο στο φύλο;

OXI

NAI

Ουδέτερο; κλπ

35

(πβλ. Harley & Ritter 2002)

Οριστικό άρθρο;

Εξερευνώντας το WALS και άλλες τυπολογικές βάσεις

Σημείο αφετηρίας ήταν το World Atlas of Language Structures (WALS; Dryer & Haspelmath 2013), το οποίο διαθέτει τα παρακάτω πλεονεκτήματα:

1. Είναι η **μεγαλύτερη και πιο πλήρης βάση δεδομένων**, η οποία καλύπτει ένα ευρύ φάσμα γραμματικών δεδομένων ('features' στο WALS) όπως αυτά εμφανίζονται σε ένα μεγάλο αριθμό γλωσσών του κόσμου.
2. Παρέχει πληροφορίες με υψηλό ποσοστό πληρότητας σε ένα δείγμα **100 γλωσσών σταθμισμένο γεωγραφικά και ιστορικά**.
3. Ο συνδυασμός μεγάλου αριθμού χαρακτηριστικών προερχόμενων από όλες τις περιοχές της σύνταξης και της μορφολογίας **ελαχιστοποιεί την επίδραση της ομοπλασίας**.
4. Επιτρέπει μεγαλύτερης τάξης μέγεθος στον **αριθμό των διαθέσιμων ταξονομικών χαρακτήρων**, απαραίτητο για τη φυλογενετική έρευνα.
5. Δεδομένα τέτοιου τύπου είναι τα μόνα τα οποία μπορούν να χρησιμοποιηθούν για την επανασύνθεση **σχέσεων μεταξύ γλωσσικών οικογενειών** (βαθείς φυλογένειες).
6. Μπορεί να **συνδυαστεί με άλλες τυπολογικές βάσεις δεδομένων**, όπως το SSWL, οδηγώντας σε περαιτέρω εμπλουτισμό των δεδομένων.

Περιορισμοί του δείγματος των 100 γλωσσών του WALS

1. Το δείγμα των 100 γλωσσών μπορεί να χρησιμοποιηθεί **μόνο για μεγάλης κλίμακας φυλογενετική έρευνα** καθώς είναι ένα αντιπροσωπευτικό δείγμα των γλωσσικών οικογενειών του κόσμου.
2. Περιλαμβάνει **λίγες γλώσσες ανά οικογένεια**. Ιδανικά, μπορεί να χρησιμοποιηθεί για να αναπαραγάγει βαθιές γενεαλογικές σχέσεις ανάμεσα σε οικογένειες. Πρακτικά, μπορεί να φανεί εν τέλει ότι επανασυνθέτει ρηχές σχέσεις ανάμεσα σε γλώσσες της ίδιας οικογένειας.
3. Πρακτικοί περιορισμοί: Πολλές τιμές που λείπουν, ανακρίβειες, ασυνέπειες και λάθη τα οποία είναι δύσκολο να ελεγχθούν και να επαληθευθούν καθώς πρόκειται και για γλώσσες που έχουν μελετηθεί σε μικρότερο βαθμό.

Μία βάση δεδομένων 60 ΙΕ γλωσσών βασισμένη στο WALS I

Τα χαρακτηριστικά του WALS δέχονται πολλαπλές τιμές και χωρίζονται στις εξής κατηγορίες:

- › φωνολογία (20), μορφολογία (12), ονοματικές κατηγορίες (29), ονοματική σύνταξη (8), ρηματικές κατηγορίες (17), σειρά όρων (56), απλές προτάσεις (26), σύνθετες προτάσεις (7), λεξικό (13).

Αποτελούν καλό σημείο εκκίνησης, επειδή μπορέσαμε να συλλέξουμε χαρακτήρες που καλύπτουν όλα τα πεδία της μορφολογίας και της σύνταξης.

1. Για να ξεπεράσουμε τους περιορισμούς του δείγματος των 100 γλωσσών, **αναπτύξαμε έναν καινούριο πίνακα με 60 ΙΕ γλώσσες**, λαμβάνοντας υπόψη χαρακτηριστικά του WALS από τις παρακάτω κατηγορίες:

- › μορφολογία (12), ονοματικές κατηγορίες (28), ονοματική σύνταξη (8), ρηματικές κατηγορίες (17), σειρά όρων (56), απλές προτάσεις (24), σύνθετες προτάσεις (7), λεξικό (4). Σε αυτό το στάδιο ένας μικρός αριθμός χαρακτηριστικών αποκλείστηκε ως μη πληροφοριακά.

Μία βάση δεδομένων 60 ΙΕ γλωσσών βασισμένη στο WALS II

2. Από αυτά τα χαρακτηριστικά, επιτεύχθηκε πληρότητα $\geq 75\%$ για τα παρακάτω:

- › μορφολογία (8),
- › ονοματικές κατηγορίες (23),
- › ονοματική σύνταξη (7),
- › ρηματικές κατηγορίες (13),
- › σειρά όρων (20),
- › απλές προτάσεις (18),
- › σύνθετες προτάσεις (1),
- › λεξικό (2).

Τα περισσότερα από τα χαρακτηριστικά που αποκλείστηκαν δεν εφαρμόζονταν στην ΙΕ (π.χ. πολλά χαρακτηριστικά από τη «σειρά όρων» ορίζονταν υπό συνθήκες που απουσιάζουν από τις ΙΕ γλώσσες) ή χαρακτηères για τους οποίους ήταν δύσκολο να βρεθούν ακριβή δεδομένα εξαιτίας ασάφειας στον ορισμό (π.χ. 22Α «Κλιτική σύνθεση στο Ρήμα»).

Μία βάση δεδομένων 60 ΙΕ γλωσσών βασισμένη στο WALS III

3. Από τα $60 \times 92 = 5520$ σημεία δεδομένων που προέκυψαν, το 30% περίπου μπορούσε να συμπληρωθεί με βάση τις τιμές που υπήρχαν στο WALS.

Για να έχουμε όσο το δυνατόν πιο αξιόπιστα δεδομένα, με βάση γραμματικές, γλωσσολογική βιβλιογραφία και τη διαίσθηση πληροφορητών:

- i. Επαληθεύσαμε την ακρίβεια των υπαρχουσών πληροφοριών, και
- ii. συμπληρώσαμε τις τιμές που έλειπαν.

Μία βάση δεδομένων 60 ΙΕ γλωσσών βασισμένη στο WALS IV

4. Στη συνέχεια μετατρέψαμε τα χαρακτηριστικά με πολλαπλές τιμές σε χαρακτήρες με δυαδικές τιμές, στην ουσία μετατρέποντας κάθε ένα από τα 5345 σύνολα τιμών (των χαρακτηριστικών που αναφέρθηκαν στο 3) νωρίτερα) σε μία ερώτηση ολικής αγνοίας για το αν η κατάσταση που περιγραφόταν από την τιμή ήταν αληθής για την υπό διερεύνηση γλώσσα. Η διαδικασία αυτή μας οδήγησε σε 16380 χαρακτήρες με δυαδικές τιμές.
5. Επίσης επισημειώσαμε τους χαρακτήρες μας με βάση την κλίμακα που αναπτύχθηκε από τους Wichmann & Holman (2009), το οποίο ομαδοποιεί τα χαρακτηριστικά του WALS και τις καταστάσεις τους σε τέσσερις τύπους σύμφωνα με τη διαχρονική σταθερότητά τους: πολύ σταθερούς, σταθερούς, ασταθείς, πολύ ασταθείς.

Μία βάση δεδομένων 60 ΙΕ γλωσσών βασισμένη στο WALS IV: Επιλογή Γλωσσών

- Σύγχρονες Γλώσσες (53)
- Ευρωπαϊκές Γλώσσες (44)
 - i. Αλβανικά
 - ii. Αρμένικα
 - iii. ΚΝΕ, Κυπριακά, 3 Ιταλιώτικες & 5 Μικρασιατικές ποικιλίες
 - iv. Ρομανικές Γλώσσες (7)
 - v. Κελτικές Γλώσσες (4)
 - vi. Βαλτο-σλαβικές Γλώσσες (12)
 - vii. Γερμανικές Γλώσσες (9)
- Ινδο-Ιρανικές Γλώσσες (9)
 - i. Ινδικές (7)
 - ii. Ιρανικές (2)

A. Παλαιότερα συστήματα (7)

- i. Κλασική Ελληνική
- ii. Ελληνικά Καινής Διαθήκης
- iii. Λατινική
- iv. Παλαιά Εκκλησιαστική Σλαβονική
- v. Γοτθικά
- vi. Παλαιά Αγγλικά
- vii. Σανσκριτικά

Τα δεδομένα του ModelGloss

Τα πρώτα δεδομένα του ModelGloss περιλαμβάνουν:

- Πίνακες γλωσσικών χαρακτήρων και των τιμών τους από The World Atlas of Language Structures (<https://wals.info>)
- Δεδομένα από το δείγμα των 100 γλωσσών του WALS sample of WALS

		Characters					
		SC98	SC99	SC100	SC101	SC102	SC103
L a n g u a g e s	hix	1	0	1	0	0	0
	hmo	1	0	1	0	1	1
	imo	0	0	0	0	1	0
	ind	1	0	1	0	0	1
	jak	1	0	1	0	0	0
	jpn	0	1	0	1	1	0
	knd	0	1	0	1	0	0
	krk	1	0	1	0	0	0
	kay	0	1	0	1	1	0
	kew	0	0	0	0	0	0

Τα δεδομένα του ModelGloss: Εμπλουτίζοντας τη βάση δεδομένων

- Πίνακα 156 χαρακτηριστικών του WALS με τιμές που συλλέχθηκαν για το ModelGloss για 60 Γλώσσες της Ινδο-Ευρωπαϊκής.

		Characters					
		37A	38A	39A	40A	41A	42A
L a n g u a g e s	Italian	1	2	3	3	2	1
	Spanish	1	2	3	3	3	1
	Catalan	1	2	3	3	3	1
	French	1	2	3	3	1	2
	EU Portug.	1	2	3	3	3	1
	Br. Portug.	1	2	3	3	3	1
	Romanian	3	2	3	3	2	1
	Latin	5	5	3	3	4	1
	Cl. Greek	1	4	3	3	3	1
	New Test.	1	4	3	3		1

Τα δεδομένα του ModelGloss: Μετατροπή σε δυαδικούς χαρακτήρες

- Από τα 156 χαρακτηριστικά του WALS, επετεύχθη πληρότητα $\geq 75\%$ για 92.
- Μετατρέψαμε τις τιμές αυτών των 92 χαρακτηριστικών σε ερωτήσεις ολικής αγνοίας που ρωτούσαν εάν η κατάσταση που περιγράφεται από την τιμή του WALS είναι αληθής ή όχι για την εκάστοτε υπό εξέταση γλώσσα.
- Αυτή η μετατροπή απέδωσε 425 δυαδικούς χαρακτήρες.

		Characters				
		37A-1	37A-2	37A-3	37A-4	37A-5
Languages	Italian	1	0	0	0	0
	Spanish	1	0	0	0	0
	Catalan	1	0	0	0	0
	French	1	0	0	0	0
	EU Portug.	1	0	0	0	0
	Br. Portug.	1	0	0	0	0
	Romanian	0	0	1	0	0
	Latin	0	0	0	0	1
	Cl. Greek	1	0	0	0	0
	New Test.	1	0	0	0	0

Μέθοδοι αναπαράστασης της εξελικτικής ιστορίας με δέντρα

- Μέθοδοι αποστάσεων (Distance-based), συγκεκριμένα φυλογενετικοί αλγόριθμοι ένωσης γειτόνων
- Μέθοδοι μέγιστης φειδωλότητας (Maximum Parsimony) βασισμένες στους χαρακτήρες
- Μπεϋζιανές φυλογενετικές μέθοδοι (Bayesian phylogenetic methods - συμπεριλαμβανομένου strict clock και relaxed clock), ακολουθώντας τους Greenhill et al. (2017)

Μέθοδοι αποστάσεων

- Από τον αρχικό πίνακα με τα χαρακτηριστικά, βρίσκουμε το μέσο όρο των διαφορών ανά δύο γλώσσες (ζευγαρωτά-pairwise)

Lang1				
Lang2	4			
Lang3	5	3		
Lang4	4	4	5	
Lang5	5	3	4	1

- Κατασκευάζουμε το δέντρο με βάση αλγόριθμους που συνυπολογίζουν τις ελάχιστες διαφορές μεταξύ των γλωσσών.
- Οι βασικότερες μέθοδοι είναι η UPGMA και η Neighbor Joining.

Μέθοδοι μέγιστης φειδωλότητας (Maximum Parsimony)

- Η παλαιότερη μέθοδος
- Χρησιμοποιείται και σε μορφολογικά χαρακτηριστικά
- Ελαχιστοποιεί τον αριθμό των αλλαγών στους χαρακτήρες
- Χρησιμοποιεί τις «πληροφοριακές θέσεις» (informative sites)
- Μια πληροφοριακή θέση είναι μια θέση που έχει τουλάχιστον δύο χαρακτήρες, καθένας από τους οποίους εμφανίζονται σε τουλάχιστον δύο OTUs.

Μοντέλα Μέγιστης Πιθανοφάνειας

- Βρίσκει τις παραμέτρους ενός μοντέλου εξέλιξης οι οποίες μεγιστοποιούν την πιθανότητα να παρατηρήσουμε τα πραγματικά δεδομένα
- Το μοντέλο περιλαμβάνει:
 - › Μοντέλο αλλαγής αλληλουχίας
 - › Τοπολογία του δέντρου
 - › Μήκος των κλαδιών

Μπεϋζιανές (Bayesian) φυλογενετικές μέθοδοι

- Συνδυάζουν μια πρότερη πληροφορία για το δέντρο (prior) με ένα μοντέλο πιθανοφάνειας των δεδομένων (likelihood model) για να δημιουργήσουν την κατανομή της posterior πιθανότητας για τα δέντρα.
- Η posterior πιθανότητα ενός δέντρου είναι η πιθανότητα το δέντρο να είναι σωστό, δεδομένης της πρότερης πιθανότητας (prior), των δεδομένων και της καταλληλότητας του μοντέλου πιθανοφάνειας.
- Πρόκειται για μεθόδους πολύπλοκες μαθηματικά και απαιτητικές σε υπολογιστική ισχύ.
- Είναι αρκετά δημοφιλείς μέθοδοι διότι δίνουν τη δυνατότητα ευελιξίας

Δέντρα της ΙΕ βασισμένα στο WALs: Κύρια αποτελέσματα

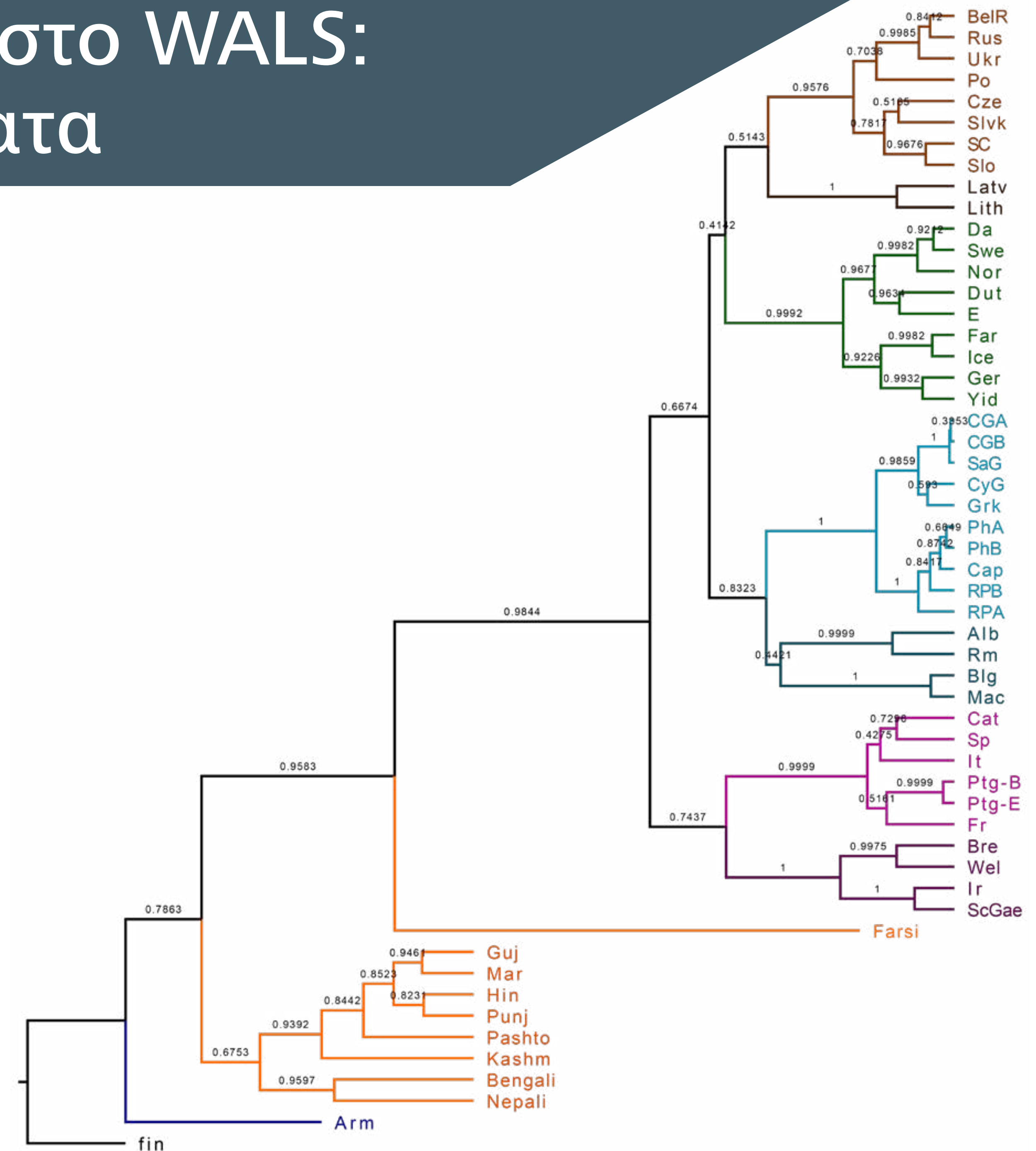
1. Σε σύγκριση με άλλες μεθόδους που αναπαριστούν την εξέλιξη με δέντρα, τα bayesian δέντρα είχαν ισχυρότερη στατιστική υποστήριξη και η τοπολογία των δέντρων ήταν πλησιέστερη στις ιστορικές και γεωγραφικές σχέσεις των γλωσσών όπως τις γνωρίζουμε.
2. Εντός των bayesian μοντέλων, δοκιμάστηκαν οι εξής εκδοχές:
 - a. μοντέλα strict και relaxed clock
 - b. δείγμα με/χωρίς τις μη-σύγχρονες γλώσσες
 - c. διαφορετικά υποσύνολα των χαρακτήρων, ανάλογα με τη σταθερότητά τους (βλ. παραπάνω βήμα 5).

Δέντρα της ΙΕ βασισμένα στο WALS: Κύρια αποτελέσματα

3. Τα περισσότερα πειράματα επανασυνθέτουν τα γένη, π.χ. Ρομανικές, Κελτικές, Ελληνικά (με τις διαλέκτους τους), Σλαβικές, Γερμανικές, Ινδο-ι-ρανικές, με τις ακόλουθες εξαιρέσεις:
 - i. έναν κλάδο που ομαδοποιεί γλώσσες του Βαλκανικού γλωσσικού δε-σμού (Sprachbund)
 - ii. έναν κλάδο που περιλαμβάνει μερικές (και κάποιες φορές και όλες) τις μη σύγχρονες γλώσσες, όποτε αυτές συμπεριλαμβάνονταν στο δείγμα γλωσσών.

Δέντρα της ΙΕ βασισμένα στο WALS: Κύρια αποτελέσματα

4. Το δέντρο που είναι πιο κοντά στην παραδεδεγμένη γνώση για την ΙΕ, προέκυψε από bayesian ανάλυση:
- ενός δείγματος που περιλάμβανε μόνο τις σύγχρονες γλώσσες
 - μοντέλο strict clock
 - πολύ σταθερούς, σταθερούς και ασταθείς χαρακτήρες (δηλαδή χωρίς τους πολύ ασταθείς χαρακτήρες).



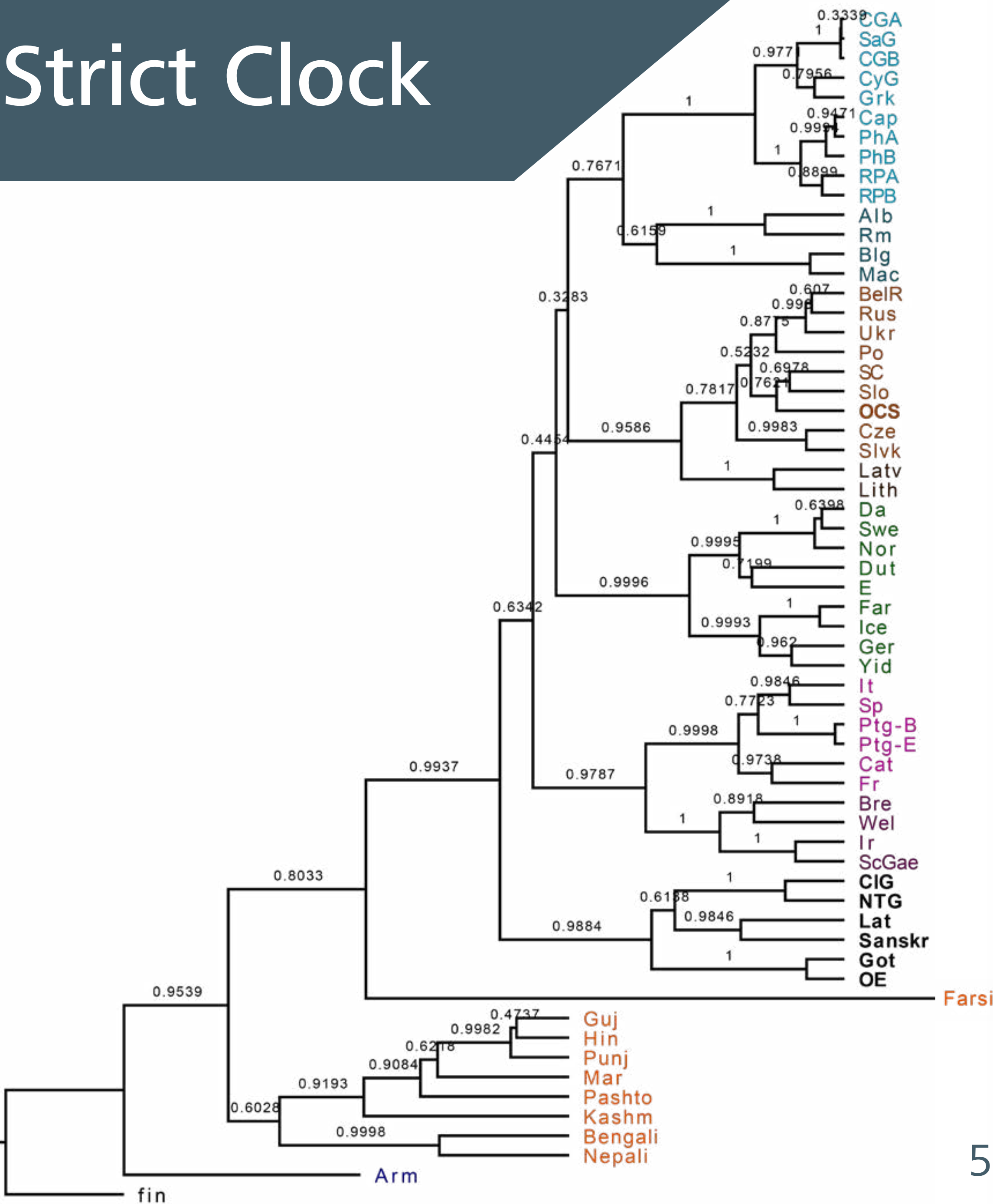
Clocks, παλαιότερες γλώσσες και ο Βαλκανικός γλωσσικός δεσμός

Η επιλογή του clock είναι σημαντική:

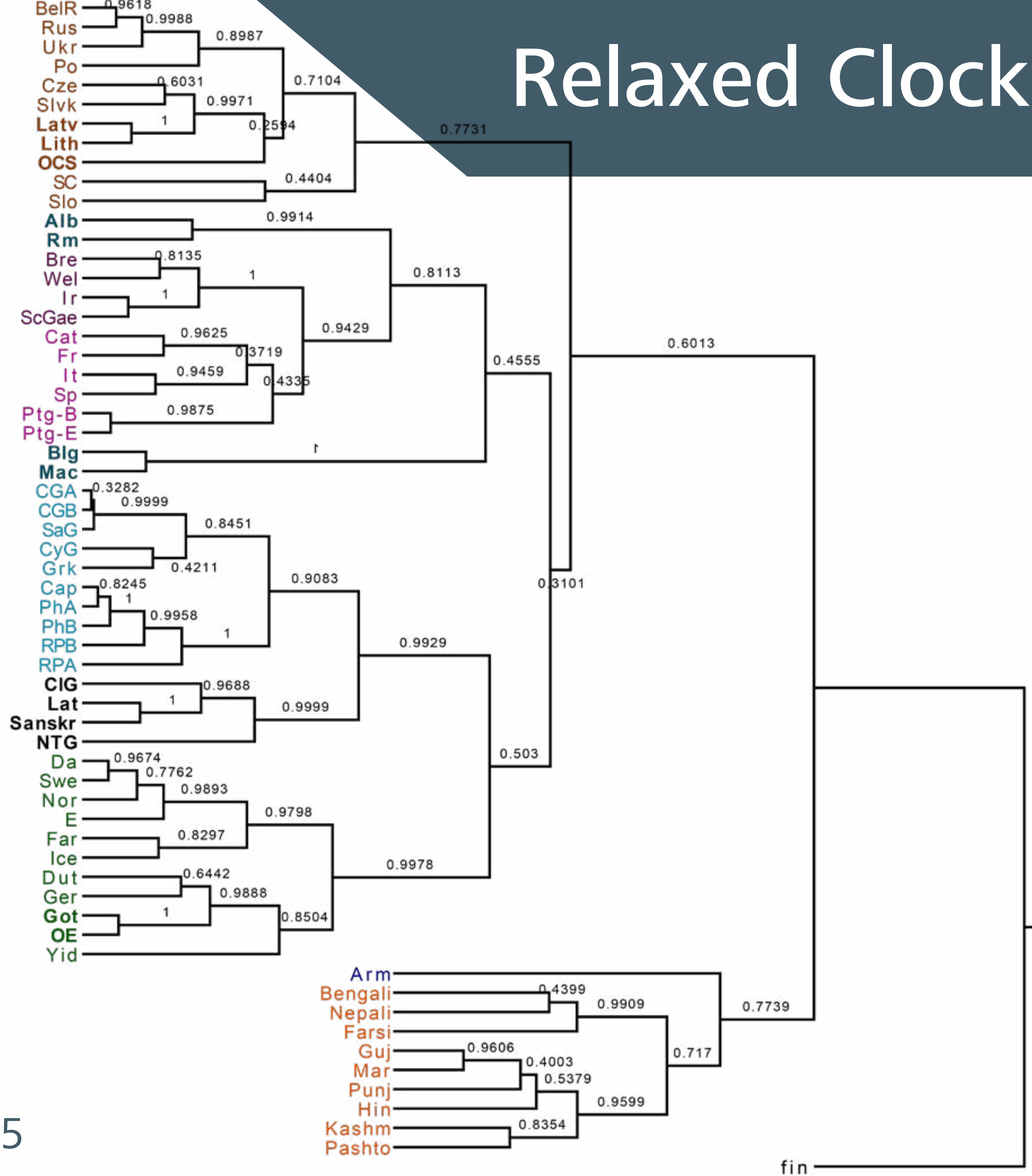
- Όταν όλοι οι χαρακτήρες συμπεριληφθούν στην ανάλυση, μόνο το relaxed clock συνδέει τις μεσαιωνικές γλώσσες με το γένος της καθεμίας (όχι όμως και τις αρχαίες).
- Από την άλλη μεριά, συσχετίσεις εξαιτίας μακροχρόνιας επαφής, όπως ο Βαλκανικός γλωσσικός δεσμός, χάνονται με το relaxed clock ενώ διατηρούνται με το αυστηρό.

Μία ιστορική ταξινόμηση των γλωσσών που ανήκουν στο Βαλκανικό γλωσσικό δεσμό δεν έχει προκύψει από τα μοντέλα που έχουν εφαρμοστεί ως τώρα.

Strict Clock



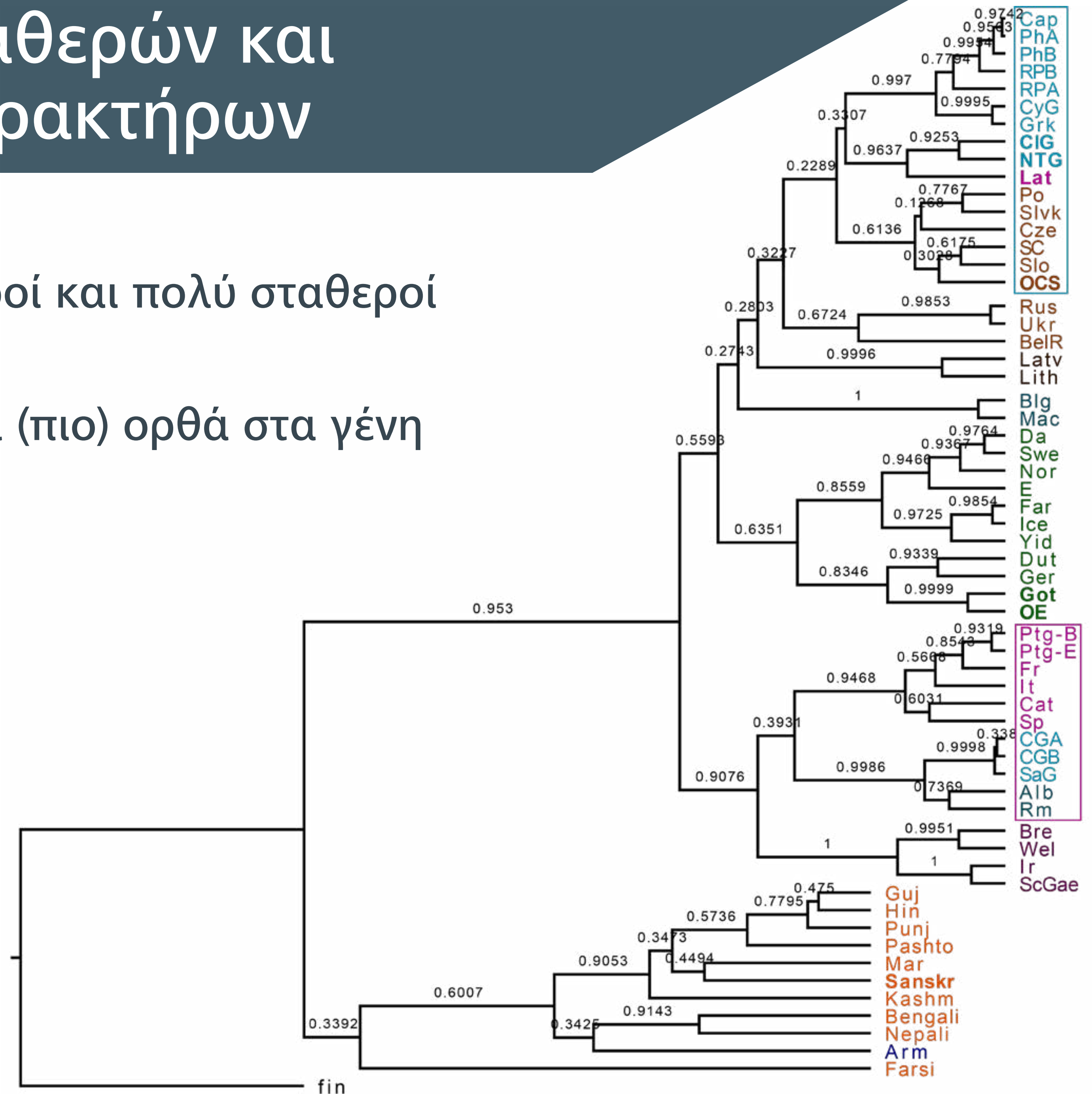
Relaxed Clock



Η επίδραση των σταθερών και πολύ σταθερών χαρακτήρων

Όταν χρησιμοποιούνται μόνο σταθεροί και πολύ σταθεροί χαρακτήρες:

- οι παλαιότερες γλώσσες συνδέονται (πιο) ορθά στα γένη τους.
- Εν τούτοις, η ενότητα οικογενειών όπως η Ελληνική, η Βαλτο-Σλαβική και η Ρομανική, και αυτών του Βαλκανικού γλωσσικού δεσμού διατηράσσεται σημαντικά.

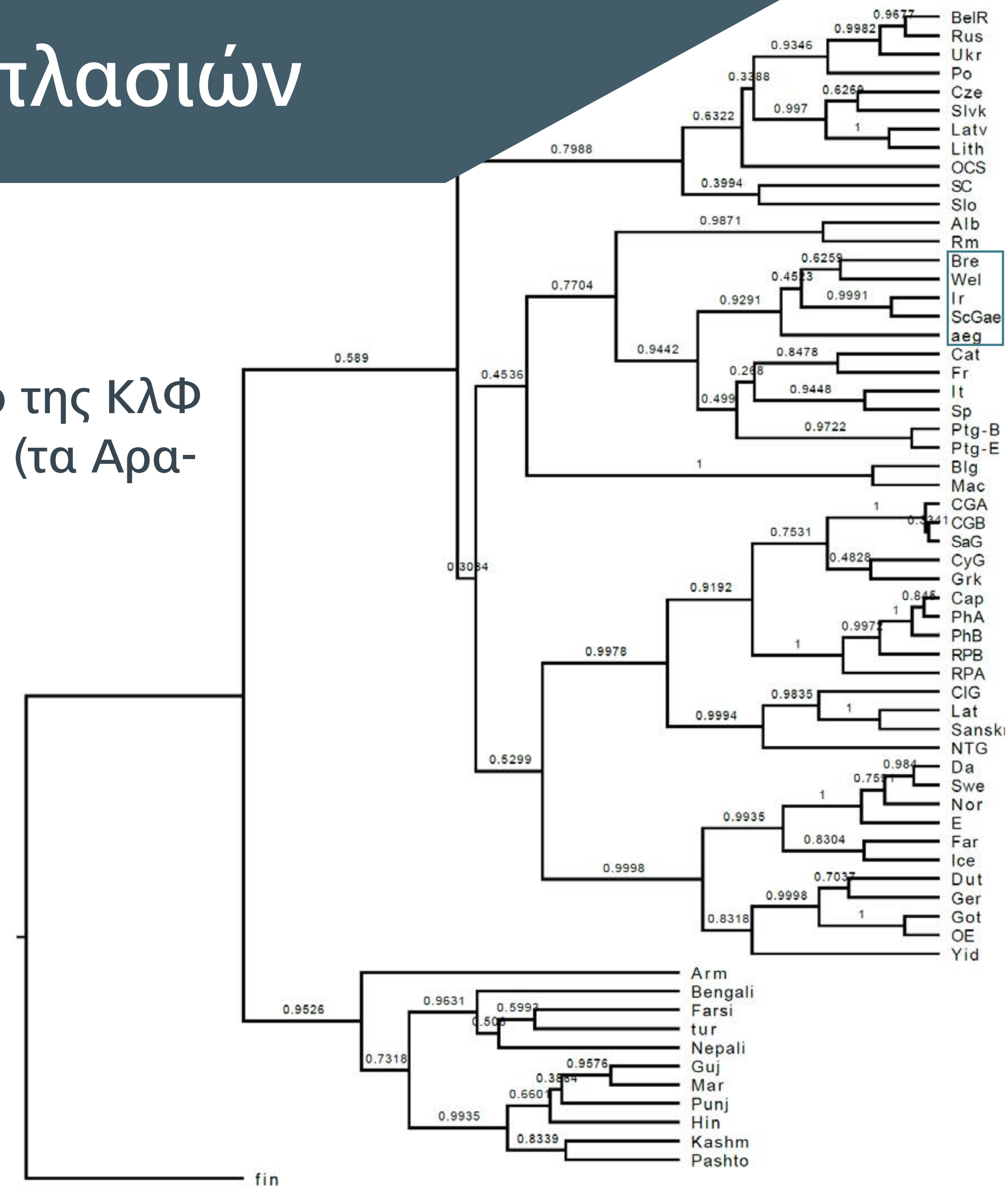


Η συνεισφορά των πολύ σταθερών χαρακτήρων

- Οι πολύ σταθεροί χαρακτήρες είναι σημαντικοί για τις βαθιές ιστορικές σχέσεις, π.χ. εξασφαλίζουν τη θέση των Φαρσί εντός της ΙΕ.
- Σε όλα τα παραπάνω δέντρα τα Φαρσί είναι εντός της ΙΕ και συνδεδεμένα με τον Ινδο-Ιρανικό κλάδο.
- Σε δέντρα που δεν συμπεριλαμβάνονται οι πολύ σταθεροί χαρακτήρες τα Φαρσί κατηγοριοποιούνται ως μη ΙΕ γλώσσα.

Η επίδραση των ομοπλασιών

- Οι γλώσσες με ισχυρές ομοιότητες στο πεδίο της ΚλΦ και της ΟΦ ομαδοποιούνται μαζί στο δέντρο (τα Αραβικά συνδέονται με τις Κελτικές γλώσσες).
- Η σύνδεση Κελτικών-Σημιτικών γλωσσών διορθώνεται όταν απουσιάζουν οι πολύ σταθεροί χαρακτήρες.
- Αυτό υποδεικνύει ότι οι πολύ σταθεροί χαρακτήρες περιλαμβάνουν πολλά ομοπλαστικά μορφο-συντακτικά χαρακτηριστικά.



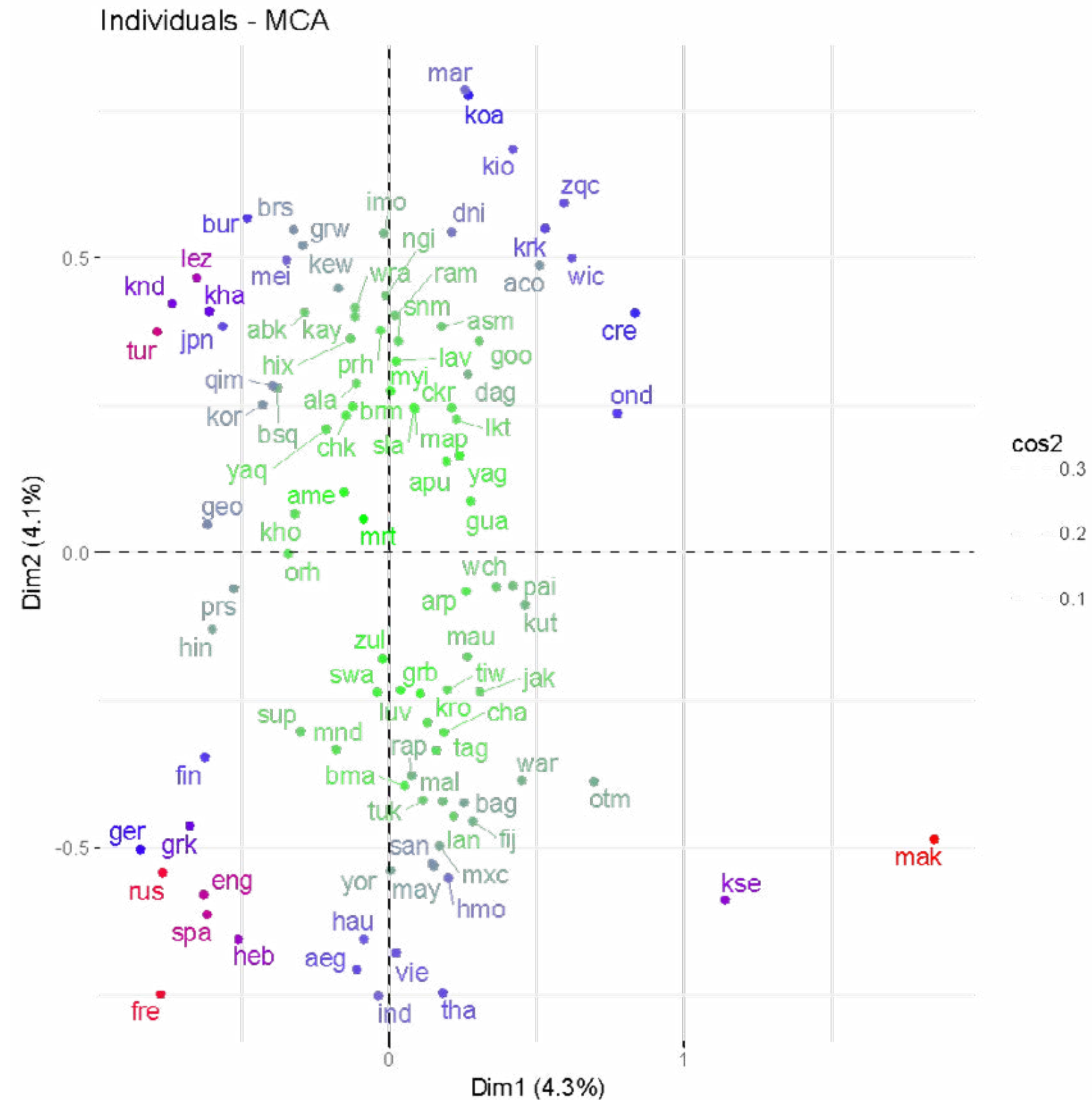
Συμπεράσματα για τα δέντρα

1. Παρά κάποιους περιορισμούς (π.χ. σύμπτυξη επιφανειακών μοτίβων τα οποία ενδεχομένως έχουν διαφορετικές δομικές αναπαραστάσεις και, κατά συνέπεια, ομοπλασίες) τα χαρακτηριστικά του WALIS παρέχουν πλούσιο ιστορικό σήμα.
2. Η κατηγοριοποίηση των χαρακτήρων με βάση τη σταθερότητα των Wichmann & Holman's (2009) έχει σημαντικές συνέπειες για την ιστορική επανασύνθεση.
3. Κανένας συνδυασμός παραμέτρων από αυτούς που ελέγξαμε δεν μπορεί να απαλείψει την επίδραση μακροχρόνιας γλωσσικής επαφής όπως για παράδειγμα οι γλώσσες του Βαλκανικού γλωσσικού δεσμού παραμένουν σε κάποιο βαθμό συνδεδεμένες. Γλώσσες όπως η Βουλγάρικη και η Ρουμάνικη δεν μπορούν να συνδεθούν στα δικά τους γένη. Αυτό θα μπορούσε να σχετίζεται και με τους χαρακτήρες του WALIS.
4. Τα μοντέλα που ελέγξαμε δεν μπορούν να διαχειριστούν ταυτόχρονα τη θέση των αρχαίων γλωσσών, τη μικροποικιλότητα και την επίδραση της γλωσσικής επαφής.

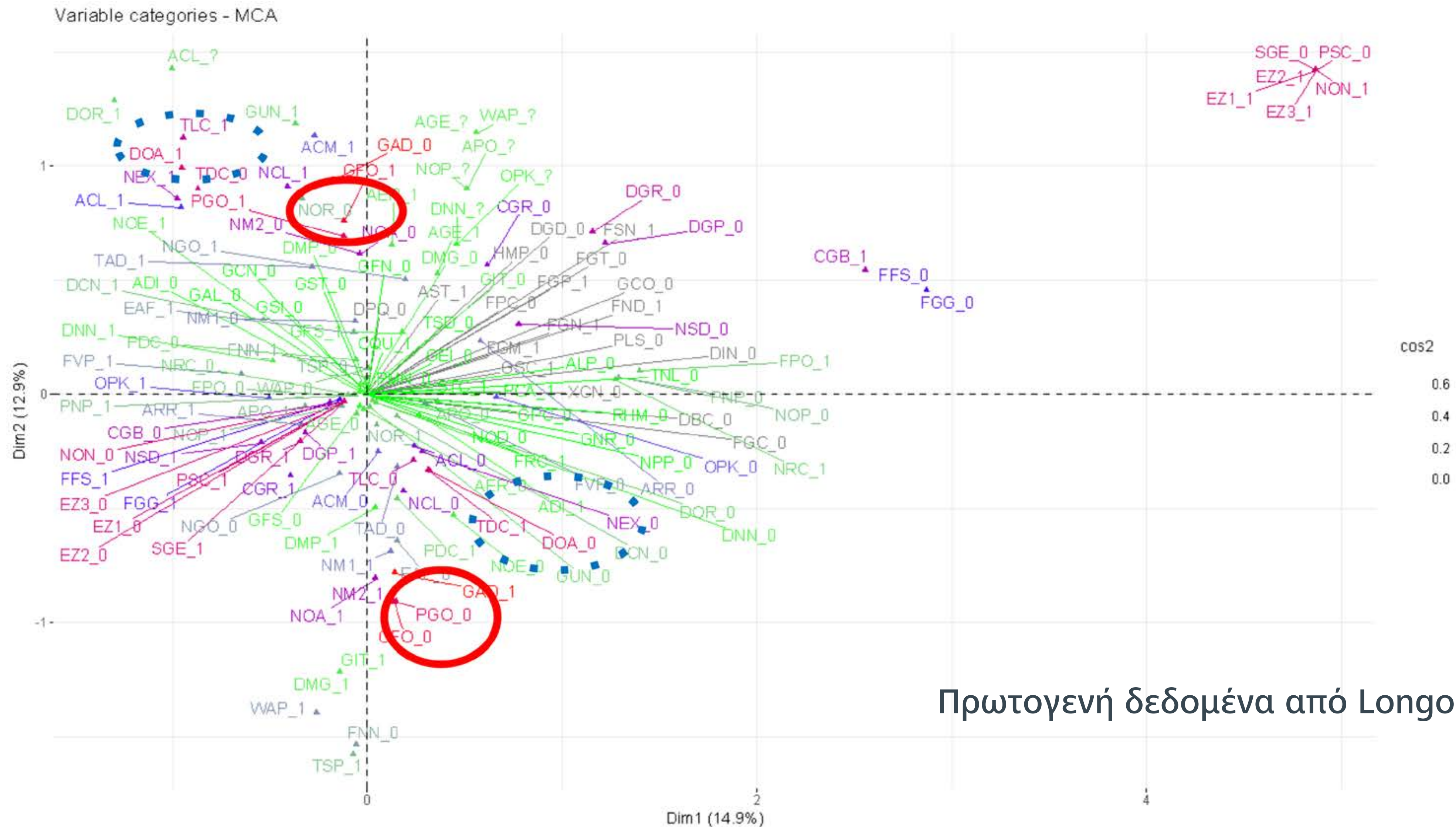
Multiple Correspondence Analysis (MCA)

- Πρόκειται για μια μέθοδο πολυμεταβλητής ανάλυσης, η οποία αναδεικνύει πρότυπα (patterns) και ομαδοποιήσεις των δεδομένων, παριστάνοντάς τα σαν σημεία σε χώρο λίγων διαστάσεων (δύο ή τριών).
- Είναι μια προέκταση της κλασικής ανάλυσης κύριων συνιστωσών (principal components analysis) αλλά με δεδομένα κατηγορικά και όχι ποσοτικά.

Multiple Correspondence Analysis (MCA) γλωσσών του WALS



Multiple Correspondence Analysis (MCA) χαρακτήρων



Πρωτογενή δεδομένα από Longobardi 2018

Εξαρτήσεις παραμέτρων: Παρατηρήσεις

- Η μέθοδος MCA επιτρέπει να απομονώσουμε αλληλοεξαρτώμενους χαρακτήρες:
 - α. SGE_0, PSC_0, EZ2_1, EZ1_1, EZ3_1, NON_1
 - β. CGB_1, FFS_0, FGGO
 - γ. PGO_0, GF0_0
 - δ. TDC_1, DOA_0
- Από αυτές τις συσχετίσεις, κάποιες είναι προφανείς (η γ) αλλά κάποιες όχι (η α, β, δ).

Εξαρτήσεις παραμέτρων: Μη προφανείς συσχετισμοί

- Το α συσχετίζει την απουσία γένους και την παρουσία συνδετικών στοιχείων (linkers) που εισάγουν προσδιορισμούς του ουσιαστικού με τη θέση του ουσιαστικού πριν από τα αριθμητικά.
- Το β συσχετίζει τον γενικευμένο ενικό (ενικό για ενικό και πληθυντικό) με την απουσία γένους και συμφωνίας στην ΟΦ (βλ. Anagnostopoulou 2016, Bayirli 2017).
- Το δ συσχετίζει τη συνύπαρξη οριστικών άρθρων και δεικτικών αντωνυμιών με τη δυνατότητα διάχυσης της οριστικότητας (βλ. Guardiano & Michelioudakis 2019).

Παράδειγμα: Συσχέτιση δ

- Τα Ελληνικά διαθέτουν δεικτικές αντωνυμίες που πρέπει να συνυπάρχουν με άρθρα και το φαινόμενο διάχυσης της οριστικότητας ενώ τα Αγγλικά όχι:
 - › Αυτό το βιβλίο / Εκείνο το βιβλίο πρβλ. * This the book (δεν λέγονται)
 - › Το κόκκινο το βιβλίο * The red the book

Σύνοψη

- Παρουσιάσαμε μία εισαγωγή στις διαδικασίες εξέλιξης των οργανισμών και των γλωσσών που νομιμοποιεί τη χρήση φυλογενετικών μεθόδων στην ιστορική επανασύνθεση/ανακατασκευή (reconstruction) των γλωσσών.
- Εξηγήσαμε πώς η σύγχρονη φυλογενετική γλωσσολογία εμπνέεται από τα εργαλεία της εξελικτικής βιολογίας.
- Παρουσιάσαμε το ModelGloss που
 - › εστιάζει στα μορφοσυντακτικά χαρακτηριστικά
 - › στοχεύει στην ανίχνευση και την αξιολόγηση του ιστορικού τους σήματος και
 - › στην κατασκευή ιεραρχιών χαρακτήρων που οδηγούν σε παραμέτρους.
- Συζητήσαμε τα πρώτα αποτελέσματα με δέντρα που στηρίζονται στους χαρακτήρες του WALIS για την Ινδοευρωπαϊκή οικογένεια.
- Τέλος, παρουσιάσαμε τη μέθοδο MCA και πώς αυτή μπορεί να χρησιμοποιηθεί για τη διερεύνηση εξαρτήσεων χαρακτήρων.



Η ερευνητική εργασία υποστηρίχτηκε από το Ελληνικό Ίδρυμα Έρευνας και Καινοτομίας (ΕΛ.ΙΔ.Ε.Κ.) στο πλαίσιο της Δράσης 1η Προκήρυξη ερευνητικών έργων ΕΛ.ΙΔ.Ε.Κ. για την ενίσχυση των μελών ΔΕΠ και Ερευνητών/τριών και την προμήθεια ερευνητικού εξοπλισμού μεγάλης αξίας (Αριθμός Έργου: HFRI-FM17-44)

Βιβλιογραφία

- Biberauer, T., & Roberts, I. (2016). Parameter typology from a diachronic perspective. *Theoretical approaches to linguistic variation*, 234, 259–291.
- Dryer, M. S., & Haspelmath, M. (2013). The world atlas of language structures online.
- Fitch, T. W. (2010). *The Evolution of Language*. Cambridge: Cambridge University Press.
- Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965), 435–439.
- Greenhill, S. J., Wu, C. H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D. (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, 114(42), E8822-E8829.
- Harley, H., & Ritter, E. (2002). Person and Number in Pronouns: A Feature-Geometric Analysis. *Language* 78, 482–526.
- Hurford, J. (1990). Nativist and functional explanations in language acquisition. In I. M. Roca (Ed.) *Logical Issues in Language Acquisition*. Dordrecht: Foris Publications, 85–136.
- Jäger, G. (2015). Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, 112(41), 12752–12757.
- Longobardi, G., & Guardiano, C. (2009). Evidence for syntax as a signal of historical relatedness. *Lingua*, 119(11), 1679–1706.
- Longobardi, G., Ceolin, A., Ecay, A., Ghirotto, S., Guardiano, C., Irimia, M. A, Michelioudakis, D., Radkevich, N., Pettener, D., Luiselli, D. & Barbujani, G. (2016). Formal linguistics as a cue to demographic history.
- Nichols, J. (1992). *Linguistic diversity in space and time*. University of Chicago Press.
- Ringe, D., Warnow, T., & Taylor, A. (2002). Indo-European and computational cladistics. *Transactions of the philological society*, 100(1), 59–129.
- Roberts, I. (2019). *Parameter hierarchies and universal grammar*. Oxford University Press, USA.
- Sarris, P. F., Ladoukakis, E. D., Panopoulos, N. J., & Scoulica, E. V. (2014). A phage tail-derived element with wide distribution among both prokaryotic domains: a comparative genomic and phylogenetic study. *Genome biology and evolution*, 6(7), 1739–1747.
- Schleicher, A. (1863). Darwinism tested by the science of language, trans. AV M. Bickers. Böhlau. Reprinted in:(1983) *Linguistics and evolutionary theory: Three essays*, ed. K. Koerner.
- Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2), 121–137.