Language evolution and dynamics of language population

Elena Maslova

Question:

- There is a lot of variation among the world's languages today (word order, morphology, etc.)
- Can we "reconstruct" any traits of the earliest human language(s) (which word order, what (if any) morphology, etc.)?

- ✓ Language “maturation” (Dahl 2002):
  - certain types of “complexity” in language can arise only in the course of “maturation”
  - ? the earliest language(s) were “young”, hence in some respects “simpler” than many modern languages (~ pidgins)
- ✓ The distribution of language types today vs. tendencies of language change
  - languages acquire nominal conjunction (Mary AND John), but appear not to “lose” it (Stassen), but there are a lot of languages without AND now (Mary came WITH John)
  - ? the earliest languages had no AND

Reconstruction of “Proto-World” word order (Newmeyer 2000):

- ✓ SOV order predominates among the world’s languages today
- ✓ The historical change OV > VO is more common than the change VO > OV
- ➢ SOV order was once much more typologically predominant than it is now
- ➢ The earliest human language had SOV word order

Model (Stochastic typology)

"Language evolution": $L_1 > L_2 > L_3 > L_4 \ldots$

Typological change: A > A > A > B …

$p = P(A, n + k \mid B, n)$, $q = P(B, n + k \mid A, n)$

"step" k ~ 40 generations (1000 years)

for $m \to \infty$ $P(L_m \in A) \to p/(p+q)$

for small m

$P(L_m \in A \mid L_1 \in A) = p/(p+q) + q(1-p-q)^m/(p+q)$

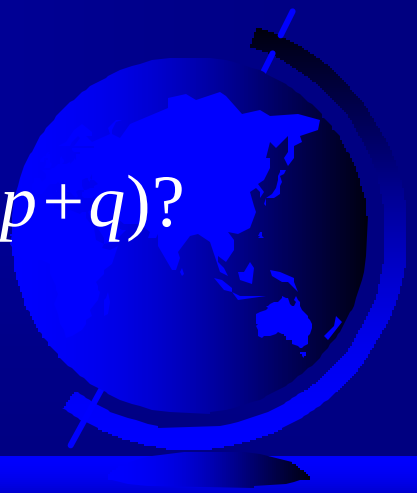
$P(L_m \in A \mid L_1 \in B) = p/(p+q) - p(1-p-q)^m/(p+q)$

Interesting questions:

- ✓ Stochastic "preferences" in language transmission: $r = p/q$
  - random or not (r ~ 0.5)?
  - unidirectional or not ($p = 0$ or $q = 0$)?
- ✓ Rate of change, or convergence time
  - $s = p+q$
  - $t \sim 1/\ln(1-s)$
- ✓ "Initial state" (L_1) -- the type of the earliest language **OR** the frequency of A in the earliest language population.

Statistical data

- $f = f_{now}(A)$ (frequency of A among the world's languages today)
- $f \approx P(L_m \in A) + \Delta f$
  - (Δf deviations due to "historical accidents")
- If the population is large, then Δf is small
  - $f \approx P(L_{10} \in A | f_{10000\ years\ ago}(A))$
- We do not know:
  - is 10 large enough to claim that $f \sim p/(p+q)$?
  - is m small enough to reconstruct L_1?



Challenge

No direct statistical data on *p* and *q*

- Estimates can be based on analysis of documented changes in a subset of languages -- for most languages, there is no written history
- They can be based on grammatical reconstructions, but they are often controversial (and sometimes based on typological considerations)

Is it possible to answer at least some of our "interesting questions"?

Proposal: “divergence rate”

- Many members of the modern language population are closely related:
  - they exist as independent languages for 1000 years or less,

  ➔ they were in the same state 1000 years ago (although we do not know in which state)
- Such pairs of languages can be of three types: (AA), (BB), and (AB), which gives us one additional “observed value”
  - $h = f$(AB) among such pairs

The last step in the history of language population

- $f \approx f^{(-1)}(1-q) + (1- f^{(-1)})p$
  - languages that have retained type A + languages that acquired type A
- $h \sim 2 f^{(-1)}q(1-q) + 2(1- f^{(-1)})p(1-p)$
  - pairs that originate from type A (one language retained type A, one changed to B) + pairs that originate from type B

What the DR can tell us (I)

- **If *f* ~ *1- f*,**
  - **not much about *r* = *p:q*, but rather strong estimates for *s* = *p*+*q*, hence for convergence time. The OV:VO distribution (explored by Newmeyer) belongs precisely to this type:**
- Estimates for *f* and *h*:

  ***f*(OV) ~ 0.53 (from Tomlin's (1986) database)**

  ***h*(OV,VO) ~ 0.11 (from the same database)**

Results for OV vs. VO: preferences in language transmission

- **$0.3 < p/q < 4$**

- If $p/q < 0.3$, then the modern frequency f would be significantly less than 0.53 even if the "initial" frequency $f^{(-10)} = 1$
- If $p/q > 4$, then the modern frequency f would be significantly higher than 0.53 even if the "initial" frequency $f^{(-10)} = 0$.

OV vs. VO convergence time

- **$0.112 < p+q < 0.118$**

- It takes ca. 24000 years for the state of a language to become independent on its initial state
- The modern distribution cannot depend on the type of “initial” language(s) (the rate of change is too high)
- It can depend on the distribution 10000 years ago.

What DR can tell us (II)

- **If *f* >> *1- f*,**
  - **fairly good estimates for *r* = *p*:*q* and *s* = *p*+*q*. An example: SO vs. OS distribution**
- Estimates for *f* and *h*:

  ***f*(SO) ~ 0.96 (from Tomlin's (1986) database)**

  ***h*(SO,OS) ~ 0.065 (from the same database)**

Results for SO vs. OS preferences in language transmission

- **$7 < p/q < 21$**

- ✓ If $p/q < 7$, then the modern frequency f would be significantly less than 0.96 even if the "initial" frequency $f^{(-10)} = 1$
- ✓ If $p/q > 21$, then all languages would have changed to A long ago even if the "initial" frequency $f^{(-10)} = 0$. NB: no unidirectionality

SO vs. OS convergence time

- $p+q > 0.264$

- It takes less than 11000 years for the state of a language to become independent on its initial state
- The modern distribution cannot depend on the type of “initial” language(s) (the rate of change is too high); it must be fairly close to the stationary distribution.
- **p:q > 15**

Results for SV vs. VS

- f(SV) ~ 0.86, h(SV,VS) ~ 0.065
- Preference $p/q > 3$
  - (NB: unidirectionality hypothesis (q = 0) cannot be rejected)
- $0.107 < p+q < 0.235$

➢ convergence time is less than 30000 and more than 10000

➢ No inferences about “initial state”; a weak (but still significant) estimate for p/q

Summary

- Results are good for typology (Greenbergian universals), i.e., its underlying assumptions are, by and large, confirmed by this analysis.
- But what about “earliest” human languages?
  - Negative answer so far: the rate of change is high enough for any dependencies on the “initial state” to be discernible now.
  - Is there a theoretical possibility to detect such dependencies if they do exist?

Perspectives

- Word order parameters are known to be relatively unstable (~ to have relatively short convergence time). It may be a dependency on the initial state exists for other typological distributions…
- This procedure will also easily detect very "recent" types, i.e., types which did not exist, say, 4000 years ago. If such a type is detected (e.g., AND-languages?), then it can be assumed that the earliest languages did not belong to this type...