

Linguistics 470

Lexicostatistics and Glottochronology



Today's topics



- Quantifying relationships between related languages
 - **Lexicostatistics**: a measure of linguistic similarity
 - **Glottochronology**: a method of dating based on lexicostatistics
 - The methods involve counting shared percentages of vocabulary and arriving at a measure of similarity and depth of relationship based on these percentages.
 - Popular in the 1950s-60s; now reviled by most linguists; still popular among non-linguists
 - Same reason as with megalocomparison
 - Numerical aspect also appeals to mathematicians, etc.—seems scientific
- We will evaluate **linguistic** aspects of these proposals, steering clear of their **statistical** aspects.



Lexicostatistics

- Measure the percentage of cognates in [basic word lists](#).
 - **Cognates** = similar words with similar meanings in two languages where the similarity is attributable to descent from a common ancestral form in an ancestral language.
 - E.g. in comparing 'head':
 - English *head* : German *Kopf* would count (*kaput)
 - English *head* : French *tête* would not count (< *testa)
- The larger the percentage of cognates, the more recently the two languages being compared are presumed to have separated.



Lexicostatistics and subgrouping

Level of Grouping	%	(time depth)
dialects	81-100	0-500
languages of 'family'	36-81	500-2500
'family' of stock	12-36	2500-5000
stock of microphylum	4-12	5000-7500
micro. of macrophylum	1-4	7500-10000
meso. of macro.	0-1	> 10000

Kroeber 1955: < 8% is statistically unreliable

calculated using the glottochronological formula, which we will get to later

Lexicostatistics: summary



- Notice that the lexicostatistical method makes an initial assumption of relatedness.
- The method typically does not distinguish **borrowings, etc.** from **genetic inheritance** and so is flawed as a method of determining genetic relationship.
- It is useful only in so far as it provides rough groupings of languages within which we can then look for more substantial evidence for relationship.

Glottochronology



- “A controversial method of assessing the temporal divergence of two languages based on changes of vocabulary (lexicostatistics), and expressed as an arithmetic formula.”
 - Concept: Swadesh 1950 et seqq.
 - Original 200-word list pruned to 100 (Swadesh 1955) to exclude culture-dependent items like animal names (fish), items dependent on climate (snow), and items that could be expressed by synonyms (woman/wife)
 - Statistical component: Lees 1953
- Analogous to the use of C14 dating of organic materials in that a “lexical half-life” is estimated and used to extrapolate to the time the two languages being compared diverged.
- **Basic assumptions:**
 - Basic/Core vocabulary: there exists a core vocabulary that is universal, is relatively culture-independent, and which is less subject to replacement. (Swadesh lists)
 - Constant rate of retention through time (81% over 1,000 years for his 200-item list of core vocabulary, 86% for his 100-word list)
 - Constant rate of loss cross-linguistically (Lees)
 - $t = \log c / 2 \log r$
 - t - time of separation
 - c - percentage of shared core vocabulary
 - r - the glottochronological constant (= 81/86%)

Glottochronology



- An example:

- In languages A and B, the percentage of shared cognates is 70%.

- $t = \log c / 2 \log r$

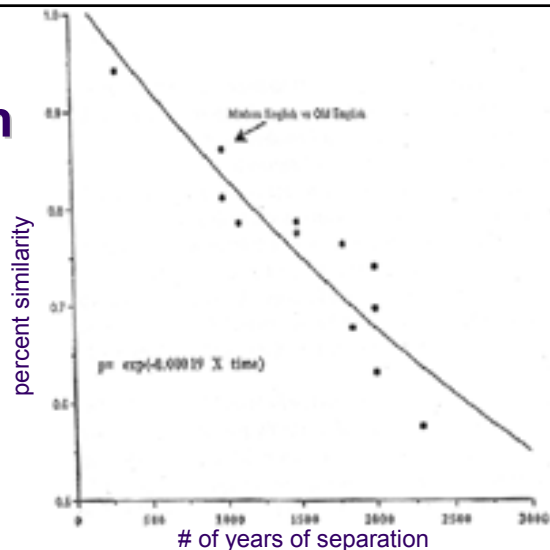
- c (% of shared core vocabulary) = 0.7

- r (glottochronological constant) = 0.86

$$t = \frac{\log 0.7}{2 \log 0.86} = \frac{-0.357}{2x - 0.151} = 1.182$$

- $t = 1.182$, i.e. 1182 years

Swadesh



- Graph depicts percentage of shared cognate words in Swadesh's fundamental vocabulary between pairs of ancestor-descendant languages (Swadesh 1952).
- Curve fitted by least squares accounts for 87% of the variance.
- Exponent of 0.00019 corresponds to a rate of approximately **20% divergence per millennium** (Swadesh's Rule).



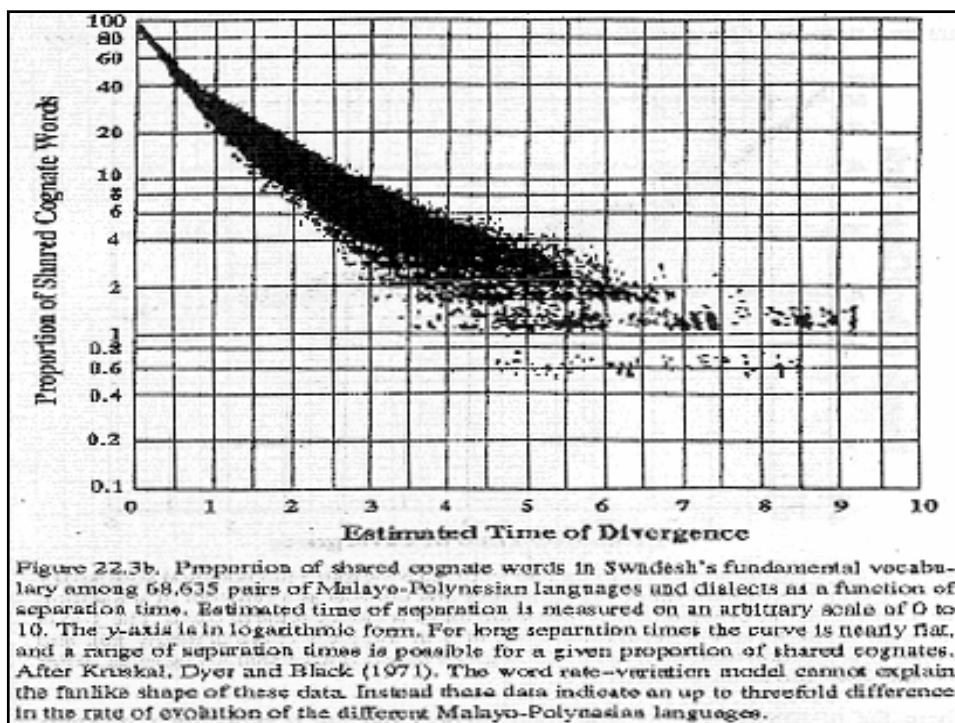
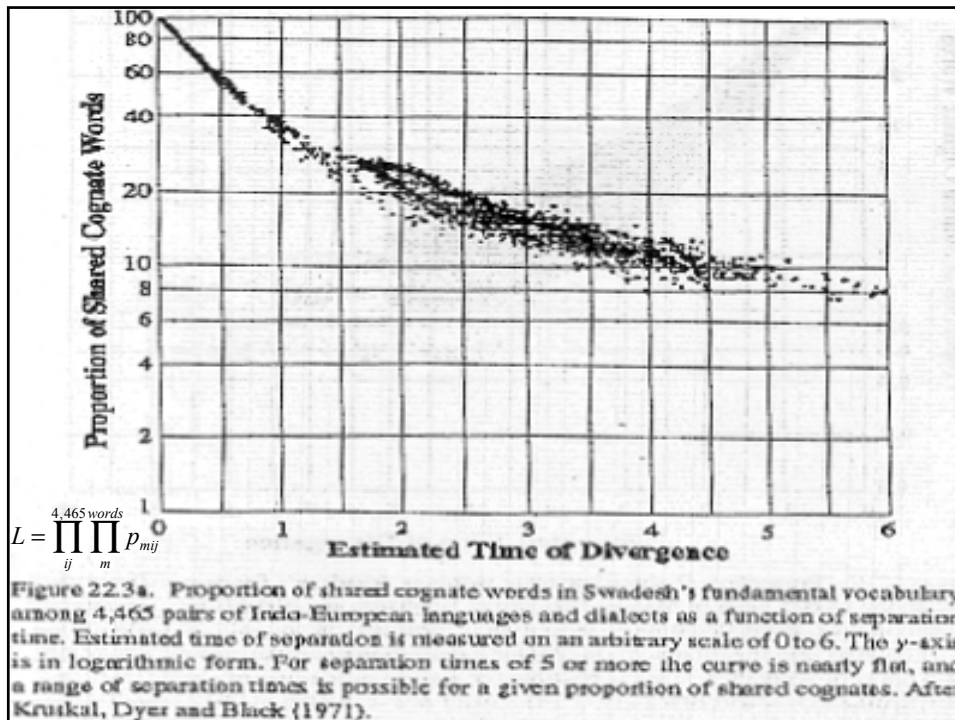
Cases where it “works”

- Hamito-Semitic (Fleming 1973)
- Chinese (Munro 1978)
- Amerind
 - Correlation with radiocarbon dating and blood groups (Stark 1973)
 - Correlation with archeological evidence for entry of Palaihnihan peoples (Atsugewi, Achumawi) into NE California (Baumhoff and Olmsted 1963)
 - The two groups are hypothesized to have separated when they entered NE California
 - 188 of Swadesh’s 200 items show up in both; 41-48 shared; ∴ 3100-3500 years since separation
 - Archeology:
 - Arrowheads, shell beads, and ornaments at Lorenzen site, near Fall River Mills, CA
 - Bottom layer at Lorenzen parallel to Central California Early Horizon materials
 - Three charcoal samples from three levels of one pit at Lorenzen carbon dated
 - 18 inches deep → c. 1452 AD (Gunther stemmed points)
 - 48 inches deep → c. 492 AD (Borax Lake points)
 - 72 inches deep → c. 1348 BC (Cascade points)
 - Major archeological discontinuity in cCA at 2000 BC, judging by radiocarbon
 - 2000 BC: original Hokan population of nCA displaced by Penutians; inhabitants of Central Valley displaced northwards; Atsugewi and Achumawi went to two different sides of the Pit River



Kruskal, Dyer & Black 1971

- 95 Indo-European languages, 4,465 pairs
- 371 Malayo-Polynesian languages, 68,635 pairs
- Initially exponential decay, then “cloud”



Problems

- **The method is highly controversial; glottochronological results are considered by many linguists to be invalid.**
 - **Basic fallacy:** a priori assumption that *all* languages change *at the same rate all the time*. This is simply not true not only regarding *different* languages but even one and the same language.
 - English and German: 75 cognates → 954 years, i.e. separated in 11th C AD!
 - Actual date: 5th C AD
 - Individual word types do not change at the same rate (e.g. numbers are more resistant)
 - Factors which affect a language's retention rate
 - Borrowing
 - Taboo
 - Strong literary tradition
 - Ethnic, national pride
 - **Lexical problems**
 - Lexicon unreliable for determining language relationships
 - Phonological and grammatical correspondences more trustworthy
 - Mistakes the dictionary for the language (common mistake among lay people)
 - Focus on lexicon ignores the core of a given language: underlying system of rules and principles
 - The problem of devising a universally valid list of basic vocabulary
 - The so-called core vocabulary is supposed to be non-cultural vocabulary, but there is no such vocabulary in a language.
 - What can or cannot have cultural content varies widely from culture to culture.
 - Two of the Tk forms (ayakta kalmak 'stand' and hayvan tirađı 'claw') are derived, leading one to question the basicness of these concepts
 - 31 of the 100 basic concepts do not appear in the 721 most basic words in Turkish; similar results for English
 - Is a list of 100 items sufficient to draw meaningful conclusions?
 - The problem of determining whether "equivalent" expressions are cognate
 - E.g. a given word may have several correspondents in the other language
 - When $n \geq 2$, use the more frequent (rules out hound : Hund; doesn't work for Tk kalp vs. gönül 'heart'; the latter is more frequent but not the citation form)
 - One can't just ask "what is the Zaza for bird?" (cf. megalocomparison)—may miss the real cognate, etc.
 - Ignores borrowing, universals, chance (cf. megalocomparison)
 - **Doesn't work for pidgins, creoles, and mixed languages** (e.g. Melanesian Pidgin; Hall 1959)

Conclusions



- Lexical comparison can be used to a very limited extent for determining closeness of relationship (but NB pidgins and creoles).
- Glottochronology probably can't be salvaged, and ignores most of the central principles of historical linguistics.

The Swadesh List (100)



all
ashes
bark
belly
big
bird
bite, to
black
blood
bone
breast
burn, to
claw
cloud
cold
come, to
die, to
dog
drink, to
dry
ear
earth
eat, to
egg

eye
fat-grease
feather
fire
fish
fly, to
foot
full
give, to
goodgreen
hair
hand
head
hear, to
heart
horn
I
kill, to
knee
know, to
leaf
lie, to
liver

long
louse
man-male
many
meat-flesh
moon
mountain
mouth
name
neck
new
night
nose
not
one
person
rain
red
road root
round
sand
say, to
see, to
seed

sit, to
skin
sleep, to
small
smoke
stand, to
star
stone
sun
swim, to
tail
that
this
thou
tongue
tooth
tree
two
walk, to
warm
water
we
what
white

who
woman
yellow

