Today's topic:

- Computational methods for subgrouping (1)

  Lexicostatistics and glottochronology

Distance-based methods for subgrouping

- Methods other than comparative reconstruction intended to find evidence for sugrouping

- Basic idea: The more closely related two languages are, the more material (words, morphemes, ...other?) they have in common

Distance-based methods for subgrouping

- *IHL*, p 137: "This method of inferring relationships does not requre reconstruction beacuse it relies on what items languages have in common, not where the commonalities come from"
  - What is this underlined phrase about? What is important in proposing subgroups based on a reconstructed protolanguage?

Lexicostatistics

- A technique sometimes used for subgrouping languages when not much data is available

- Basic assumptions:
  a. Certain parts of the lexicon are less likely to undergo lexical replacement than others (**basic** or **core** vocabulary)
  b. The rate of replacement of the core vocabulary is roughly stable, and is the same for all languages and all time periods

Core vocabulary

Claim: Certain parts of the lexicon are less likely to undergo lexical replacement than others

- Core vocabulary is claimed to be the same for all languages
  - pronouns, numerals, body parts, geographic features, basic states, ...
- A supporting example
  - English vocabulary includes ~50% borrowed forms, many from French
  - English core vocabulary has about 6% borrowings from French

Core vocabulary

Claim: The rate of replacement of the basic vocabulary is roughly stable, and is the same for all languages and all time periods

- Under these assumptions, we can determine which languages in a group **diverged recently** and which **diverged long ago** by computing the % shared core vocabulary

- Example: IHL Dataset 10

Core vocabulary — problems

- How long should the list be?
  - Longer list = more data = more accurate?
  - What if a longer list starts including ‘non-core’ vocabulary?

  → Morris Swadesh’s 200-word list often used

Core vocabulary — problems

- What semantic domains count as basic vocabulary?
  - Need to be universal — ‘brother’? ‘snow’??
  - Problem if two ‘basic’ concepts are sometimes expressed with the same form — ‘foot’, ‘leg’

Core vocabulary — problems

- The whole enterprise depends on knowing % cognate forms in a vocabulary list
- How well can we determine this?
  - Without knowledge of systematic sound changes and protolanguage forms, **we are just guessing** ("inspection method")
  - Two different linguists might come up with different % cognate for the same language
  - Some types of sound change are more likely to obscure relationships than others

Glottochronology

- Take lexicostatistical methods and assume a particular **numerical** rate of change in the basic vocabulary
  - On these assumptions, you can hypothesize a time depth for **when** two languages diverged

Glottochronology — problems

- Glottochronology crucially depends on the assumption that core vocabulary changes at a constant rate in all languages and in all time periods

- Can cultural factors influence rate of borrowing?
  - taboo / amelioration–pejoration
  - strong cultural/trade influence and number systems

Glottochronology — problems

- Suppose all languages do replace their core vocabulary at the rate of 20% every 1000 years
  - How similar will languages A and B be after 1000 years?
  - How about after 2000 years? What does this answer depend on?

Some implications

- If you assume that two languages are related, but you are wrong, and you start trying to do comparative reconstruction, what will happen?

- What if you tried this with lexicostatistical methods?

Some implications

- Glottochronology is not generally accepted by practicing historical linguists today

- Lexicostatistics can be useful as a ***first approximation*** of subgrouping in a group of languages about which not much is known