Statistical Linguistics

Amanda Llewellyn

November 14, 2013

Contents

Motivations

Why do we care about lexicostatistics?

Lexicostatistics

Lexicostatistics attempts to use numeric analysis to augment or discover similarities between languages.

- Analyze data from separate languages
- Find regular correspondences
- Hypothesize a link between languages
- Test relation for statistical significance

Word Selection

How can we standardize comparisons? Why should we?

- Standard meanings
- Potential bias in synonyms
- Choose relatively static meanings
- Cognates in meaning rather than sound
- Swadesh's 100- and 200-word lists

Glottochronology

Glottochronology attempts to extend the use of lexicostatistics to date the separation of languages. We'll examine the basic equation and the assumptions necessary for glottochronology.

- Identify two languages whose separation you want to date
- Gather data
- Calculate percentage of cognates
- Plug and chug!
- Output in kiloyears

Assumptions

- Rate of change assumptions
    - Constant across time
    - Constant across language and family
    - Modification allows for family-specific parameter
- Equation:

$$t = \frac{log(c)}{2*log(r)}$$

- t is the time in thousands of years
- c is the proportion of cognates
- r is the glottochronological constant

Example

Here, we examine several attempts at glottochronology (Trask):

- Take $r = .8$
- Consider German to English: $c = .59$
- $t = \frac{log(.59)}{2*log(.8)} = 1.18$
- Consider French to Italian: $c = .83$
- $t = \frac{log(.83)}{2*log(.8)} = .42$

Establishment of Relatedness

- Applying lexicostatistics or glottochronology to unrelated languages- bad practice
- Cognates determined by sound similarities instead of regular correspondences
- Shows bizarre, unsupported equivalences
- Basque to Berber, etc. (Trask)
- Validation preferred

Significance

Even unrelated languages can have nonzero percentages of cognates. How do we set a line which separates random chance from significant relatedness?

This requires a method of calculating an expected percentages of cognates.

- Chance resemblances?
- Mass comparison- sketchy
- Use Shift Test

More Objections

- List bias
- Only considers meanings
- Fails to integrate syntax
- Misses alternations

The End