Genetical language classification

Part 2: statistical methods
Time depth of historical reconstruction

• Three possible positions:
  – General limit for time depth that can be reached by the methods of historical linguistics. *traditional position*
  – Limits exist, but depend on available information and methods
  – No strict limit (apart from the limit set by biological evolution of human language faculty). *Greenberg, Ruhlen*
Arguments for the pessimistic position

- Reliable proof for genetic relatedness of two languages:
  - Reconstruction of common ancestor language
  - Reconstruction of the diachronic processes from the common ancestor to the languages under discussion
- Requires identification of *cognates*
- Language change obliterates similarities
- Also, cognates may be due to borrowing
- No genetic relatedness can be proved beyond 10,000 years
Arguments for the pessimistic position

Hock & Joseph (1996):

Let us pursue this issue a little further by taking a closer look at the relationship between Modern Hindi and English – pretending that we do not yet know that they are related, and trying to establish their relationship by vocabulary comparison. This is actually more difficult than it appears. It is all too easy to be influenced by one’s knowledge of the historical relationship between the two languages and therefore to notice the genuine cognates, or even to underestimate the effects of linguistic change on the recognizability of genuine cognates.
Arguments for the pessimistic position

Hock & Joseph (1996):

Clearly, one correspondence is not enough; nor are twenty. And just as clearly, a thousand correspondences with systematic recurrences of phonetic similarities and differences would be fairly persuasive. Are 500 enough, then? And if not, are 501 sufficient? Nobody can give a satisfactory answer to these questions. And this is no doubt the reason that linguists may disagree over whether a particular proposed genetic relationship is sufficiently supported or not.
Word lists

- Methods of classical comparative historical linguistics have probably reached their limit
- Alternative method: usage of **word lists**
- Identification of phonetic similarities in the vocabulary of different languages => measure of relatedness of languages
- No attempt is made to reconstruct the common ancestor language
Word lists

- Compilation of concept lists – universal basic vocabulary that is supposedly present in all languages
- Translation of this list into all languages under investigation
- For every translation pair it is tested whether there is relatedness/similarity
- Relatedness corresponds to percentage of similar words
Word lists

• Morris Swadesh (1909-1967)
  – American linguist
  – Studies a.o. The genetic classification of native American languages
  – Pioneer of lexicostatistics and glottochronology
  – Compiled the so-called Swadesh-Liste of 207 concepts that occur in all languages/cultures:
    http://www.christianlehmann.eu/fundus/Swadesh_list.html
## Lexicostatistics

- Relatedness or chance?
  *(Quelle: Maiwald & Willeke)*

<table>
<thead>
<tr>
<th>German</th>
<th>Latin</th>
<th>(engl.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herz</td>
<td>cord-</td>
<td>heart</td>
</tr>
<tr>
<td>Horn</td>
<td>cornu</td>
<td>horn</td>
</tr>
<tr>
<td>Hund</td>
<td>canis</td>
<td>dog</td>
</tr>
<tr>
<td>hundert</td>
<td>centum</td>
<td>hundred</td>
</tr>
</tbody>
</table>
## Lexicostatistics

- Relatedness or chance?

<table>
<thead>
<tr>
<th>English</th>
<th>Hawaiian</th>
</tr>
</thead>
<tbody>
<tr>
<td>sew</td>
<td>humu</td>
</tr>
<tr>
<td>smell</td>
<td>honi</td>
</tr>
<tr>
<td>snow</td>
<td>hau kea</td>
</tr>
<tr>
<td>stab</td>
<td>hou</td>
</tr>
<tr>
<td>star</td>
<td>hoku</td>
</tr>
<tr>
<td>swell</td>
<td>ho'opehu</td>
</tr>
</tbody>
</table>
Glottochronology

• Basic assumptions:
  – Every language constantly renews its vocabulary
  – This process takes place continuously
  – Vocabulary of a language has a „half-life“
  – This half-life is approximately constant across languages
  – Comparison of word lists allows reconstruction when languages have split
Glottochronology

• Calibration on the basis of comparison English-Spanish
• Estimate: after 1,000 year, 81% of the 200-Swadesh list are preserved
• Formula: \[ t = \frac{\ln c}{\ln r} \]
  - \( t \): time depth (time since proto-language)
  - \( c \): percentage of common basic vocabulary (0 < c < 1)
  - \( r \): glottochronological constant. (81%)
Glottochronology

• Advantages:
  – Applicable to any language pair
  – Little analytical effort
  – Supplies information about distant genetic relations
  – Supplies information about time depth, not just over relatedness
Glottochronology

- Disadvantages
  - Linguistic relationship is mirrored also (or even mainly) in grammar rather than in vocabulary
  - Swadesh list is not universal (some languages lack words for *if*, *five*, *smell* etc.)
  - Identification of cognates is questionable if the history of the languages are not known
  - Half-time is not constant, but depends on several factors like intensity of language contact, taboo, literary tradition, national pride, ...
Summary

• „classical“ lexicostatistics & glottochronology
  – Interesting approach
  – Based on questionable background assumptions though
  – Insufficient mathematical foundation to correctly assess the role of chance => statistics
Joseph Greenberg (1915 - 2001)

- One of the most important linguists of the 20th century
- Founder of linguistic typology
- Pathbreaking investigations to linguistic universals
- Classification of African and American languages (controversial)
Mass lexical comparison

• Greenberg:
  – Statistical reliability of lexicostatistics can be improved if the number of data points are increased
  – Rather than comparing two languages, evaluation of the basic vocabulary of entire language groups
Mass lexical comparison

Eine Klassifikationsübung nach der vergleichenden Methode à la Merritt Ruhlen:

<table>
<thead>
<tr>
<th>Sprache</th>
<th>zwei</th>
<th>drei</th>
<th>ich</th>
<th>du</th>
<th>wer?</th>
<th>nicht</th>
<th>Mutter</th>
<th>Vater</th>
<th>Zahn</th>
<th>Herz</th>
<th>Fuß</th>
<th>Maus</th>
<th>er trägt</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ñ2n-</td>
<td>-ni</td>
<td>-ka</td>
<td>man</td>
<td>lā</td>
<td>?umm-</td>
<td>abū</td>
<td>sinn</td>
<td>lubb</td>
<td>riß-</td>
<td>fār</td>
<td>yahmil-</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>sēn-</td>
<td>-ni</td>
<td>-ka</td>
<td>mi</td>
<td>lō</td>
<td>?em</td>
<td>aβ</td>
<td>šen</td>
<td>leß</td>
<td>regel</td>
<td>ṣakbōr</td>
<td>nošeh</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>duvā</td>
<td>trāyas</td>
<td>mām</td>
<td>tūvām</td>
<td>kās</td>
<td>nā</td>
<td>mātār</td>
<td>pitār-</td>
<td>dant-</td>
<td>hṛd-</td>
<td>pād</td>
<td>muṣ-</td>
<td>bhārati</td>
</tr>
<tr>
<td>D</td>
<td>duvā</td>
<td>ṭrāyō</td>
<td>mām</td>
<td>tūvōm</td>
<td>čiš</td>
<td>naē-</td>
<td>mātār-</td>
<td>ĺitar-</td>
<td>dantan-</td>
<td>zārād</td>
<td>pādīya</td>
<td>baraiti</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>duo</td>
<td>treīs</td>
<td>eme</td>
<td>sū</td>
<td>tīs</td>
<td>ou(k)</td>
<td>mātēr</td>
<td>pāter</td>
<td>odōn</td>
<td>kardīā</td>
<td>pod-</td>
<td>mūs</td>
<td>phērei</td>
</tr>
<tr>
<td>F</td>
<td>duo</td>
<td>trēs</td>
<td>mē</td>
<td>tū</td>
<td>kwis</td>
<td>ne-</td>
<td>mātēr</td>
<td>pāter</td>
<td>dent-</td>
<td>kord-</td>
<td>ped-</td>
<td>mūs</td>
<td>fert</td>
</tr>
<tr>
<td>G</td>
<td>twai</td>
<td>ṭreis</td>
<td>mik</td>
<td>ūu</td>
<td>hwās</td>
<td>ni</td>
<td>aīθēi</td>
<td>fađar</td>
<td>tunūs</td>
<td>haīrtō</td>
<td>fōt</td>
<td>baīrīθ</td>
<td>berid</td>
</tr>
<tr>
<td>H</td>
<td>dó</td>
<td>trī</td>
<td>-m</td>
<td>tū</td>
<td>kīa</td>
<td>ni-</td>
<td>māθīr</td>
<td>aθīr</td>
<td>dēt</td>
<td>kröde</td>
<td>traig</td>
<td>lux</td>
<td>sīčan</td>
</tr>
<tr>
<td>I</td>
<td>iki</td>
<td>ūč</td>
<td>ben-i</td>
<td>sen</td>
<td>kīm</td>
<td>deyīl</td>
<td>anne</td>
<td>bāba</td>
<td>diš</td>
<td>kalp</td>
<td>ayak</td>
<td>tašīyor</td>
<td></td>
</tr>
</tbody>
</table>
### Mass lexical comparison

Eine Klassifikationsübung nach der vergleichenden Methode à la Merritt Ruhlen:

<table>
<thead>
<tr>
<th>Sprache</th>
<th>zwei</th>
<th>drei</th>
<th>ich</th>
<th>du</th>
<th>wer?</th>
<th>nicht</th>
<th>Mutter</th>
<th>Vater</th>
<th>Zahn</th>
<th>Herz</th>
<th>Fuß</th>
<th>Maus</th>
<th>er trägt</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>?iθn-</td>
<td>θalāθ-</td>
<td>-ni</td>
<td>-ka</td>
<td>man</td>
<td>la</td>
<td>?umm-</td>
<td>abū</td>
<td>sinn</td>
<td>lubb</td>
<td>rijl-</td>
<td>fär</td>
<td>yahmîl-</td>
</tr>
<tr>
<td>B</td>
<td>ūn-</td>
<td>šaloš-</td>
<td>-ni</td>
<td>-ka</td>
<td>mi</td>
<td>lo</td>
<td>?em</td>
<td>aβ</td>
<td>šen</td>
<td>leβ</td>
<td>regel</td>
<td>Šaqbør</td>
<td>nošeh</td>
</tr>
<tr>
<td>C</td>
<td>duvā</td>
<td>tráyas</td>
<td>mám</td>
<td>tuvām</td>
<td>kás</td>
<td>ná</td>
<td>mātār</td>
<td>pitār-</td>
<td>dant-</td>
<td>ḥrd-</td>
<td>pád</td>
<td>muš-</td>
<td>bháratí</td>
</tr>
<tr>
<td>D</td>
<td>duva</td>
<td>ṭr̥y̥ō</td>
<td>mám</td>
<td>tuvēm</td>
<td>čiš</td>
<td>naē-</td>
<td>mā.tar-</td>
<td>pitar-</td>
<td>dant-</td>
<td>zārād</td>
<td>paidīya</td>
<td>mūs</td>
<td>phéreí</td>
</tr>
<tr>
<td>E</td>
<td>duo</td>
<td>treîs</td>
<td>eme</td>
<td>sù</td>
<td>tīs</td>
<td>ou(k)</td>
<td>māter</td>
<td>pater</td>
<td>odōn</td>
<td>kardīa</td>
<td>pod-</td>
<td>mūs</td>
<td>fert</td>
</tr>
<tr>
<td>F</td>
<td>duo</td>
<td>trēs</td>
<td>mē</td>
<td>tū</td>
<td>kwis</td>
<td>ne-</td>
<td>māter</td>
<td>pater</td>
<td>dent-</td>
<td>kord-</td>
<td>ped-</td>
<td>mūs</td>
<td>bafrīth</td>
</tr>
<tr>
<td>G</td>
<td>twai</td>
<td>ṭreis</td>
<td>mik</td>
<td>ũu</td>
<td>hwās</td>
<td>ni</td>
<td>aiθeĩ</td>
<td>fasār</td>
<td>tunb̥ūs</td>
<td>hairtō</td>
<td>föt</td>
<td>lux</td>
<td>berid</td>
</tr>
<tr>
<td>H</td>
<td>dó</td>
<td>trí</td>
<td>-m</td>
<td>tū</td>
<td>kā</td>
<td>nī-</td>
<td>máθir</td>
<td>aθîr</td>
<td>dēṭ</td>
<td>kride</td>
<td>traig</td>
<td>sičan</td>
<td>tašiyor</td>
</tr>
<tr>
<td>I</td>
<td>iki</td>
<td>üč</td>
<td>ben-i</td>
<td>sen</td>
<td>kim</td>
<td>deyil</td>
<td>anne</td>
<td>baba</td>
<td>diš</td>
<td>kalp</td>
<td>ayak</td>
<td>sičan</td>
<td></td>
</tr>
</tbody>
</table>

# Mass lexical comparison

## Multilateraler Sprachenvergleich

Schlichtes Vergleichen einiger Allerweltswörter erheilt bereits die Verwandtschaftsverhältnisse unter den Sprachfamilien Indoeuropäisch (mit den Zweigen Germanisch, Romanisch und Slawisch) sowie Uralisch-Jukagirisch und Baskisch.

<table>
<thead>
<tr>
<th>Sprachfamilie</th>
<th>Sprache</th>
<th>eins</th>
<th>zwei</th>
<th>drei</th>
<th>Kopf</th>
<th>Auge</th>
<th>Nase</th>
<th>Mund</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germanisch</td>
<td>Schwedisch</td>
<td>en</td>
<td>tvo</td>
<td>tre</td>
<td>hyvud</td>
<td>òga</td>
<td>næsa</td>
<td>mun</td>
</tr>
<tr>
<td></td>
<td>Niederländisch</td>
<td>èn</td>
<td>tvè</td>
<td>dří</td>
<td>hòft</td>
<td>òx</td>
<td>nòs</td>
<td>mont</td>
</tr>
<tr>
<td></td>
<td>Englisch</td>
<td>wèn</td>
<td>tū</td>
<td>θī</td>
<td>hèd</td>
<td>ai</td>
<td>nouz</td>
<td>mauø</td>
</tr>
<tr>
<td></td>
<td>Deutsch</td>
<td>ains</td>
<td>tsvai</td>
<td>drai</td>
<td>kópf</td>
<td>auge</td>
<td>nāzë</td>
<td>munt</td>
</tr>
<tr>
<td>Romanisch</td>
<td>Französisch</td>
<td>òë/yn</td>
<td>òø</td>
<td>trwà</td>
<td>òt</td>
<td>òej</td>
<td>ne</td>
<td>buš</td>
</tr>
<tr>
<td></td>
<td>Italienisch</td>
<td>uno</td>
<td>due</td>
<td>tre</td>
<td>testà</td>
<td>òkò</td>
<td>naso</td>
<td>boka</td>
</tr>
<tr>
<td></td>
<td>Spanisch</td>
<td>uno</td>
<td>dos</td>
<td>tres</td>
<td>kabesa</td>
<td>xo</td>
<td>naso</td>
<td>boka</td>
</tr>
<tr>
<td></td>
<td>Rumänisch</td>
<td>un</td>
<td>doi</td>
<td>trei</td>
<td>kap</td>
<td>oki</td>
<td>nas</td>
<td>gure</td>
</tr>
<tr>
<td>Slawisch</td>
<td>Polnisch</td>
<td>jeden</td>
<td>dva</td>
<td>tri</td>
<td>gwòva</td>
<td>oko</td>
<td>nos</td>
<td>usta</td>
</tr>
<tr>
<td></td>
<td>Russisch</td>
<td>adin</td>
<td>dva</td>
<td>tri</td>
<td>galàva</td>
<td>oko</td>
<td>nos</td>
<td>rot</td>
</tr>
<tr>
<td></td>
<td>Bulgarisch</td>
<td>edin</td>
<td>dva</td>
<td>tri</td>
<td>glàva</td>
<td>oko</td>
<td>nos</td>
<td>usta</td>
</tr>
<tr>
<td>Uralisch-Jukagirisch</td>
<td>Finnisch</td>
<td>yksi</td>
<td>kaksi</td>
<td>kolme</td>
<td>pæe</td>
<td>silmæ</td>
<td>nenæ</td>
<td>sū</td>
</tr>
<tr>
<td></td>
<td>Estnisch</td>
<td>yks</td>
<td>kaks</td>
<td>kolm</td>
<td>pea</td>
<td>silm</td>
<td>nina</td>
<td>sū</td>
</tr>
<tr>
<td>Baskisch</td>
<td>Baskisch</td>
<td>bat</td>
<td>bi</td>
<td>hìyr</td>
<td>byry</td>
<td>begi</td>
<td>sydyr</td>
<td>aho</td>
</tr>
</tbody>
</table>
Africa

- Greenberg 1963 „The languages of Africa“
  - Only four language families in Africa
    - Afroasiatic (replaces traditional „Hamito-Semitic“)
    - Niger-Kongo
    - Nilosaharian
    - Khoisan
  - Not uncontroversial, but largely accepted
1971: „The Indo-Pacific Hypothesis“
- Indo-Pacific macro family
- Comprises Papua languages, andamanian and tasmanian languages
Nowadays generally rejected
America

- 1987: „Language in the Americas“
- Three macro-families:
  - Eskimo-Aleut
  - Na-Dené
  - Amerind
- Latter class is heavily contested
- Standard wisdom: Greenberg's Amerind consists of about 200 families
Eurasien

  – Macro-family „Eurasiatic“
  – sub-families:
    • Indoeuropean
    • uralic languages
    • Altaic (Turkic, Mongolian, Tungusian, Korean, Japanese)
    • Eskimo-Aleut
    • Various isolated languages (like Etruskian)
  – Also highly controversial
Ruhlen

- Merritt Ruhlen
  - Student of Greenberg
  - Even more radical application of Greenberg's method of genetic classification
  - Hypothesis that partial reconstruction of „Proto-World“ is possible – the original language of humankind
  - Generally rejected by experts
Criticism of Greenberg/Ruhlen

- Fundamental objections against the usage of word lists
  - Not culture-independent
  - Assumptions of „universal“ concepts is questionable

- Specific objections against Greenberg's method
  - Very loose interpretation of „semantic correspondence“
  - No serious statistical evaluation
  - Boë et al. 2003: Ruhlen's reconstruction of Proto-World is based on statistically non-significant data (similarities could be due to chance)
  - Greenberg repeatedly reached valuable results, but so far nobody else managed to apply his method successfully => success perhaps more due to his extremely good intuition than to the validity of his method

Intuition als aufgrund einer validen Methode
The genetic family tree of humankind

- Luigi Luca Cavalli-Sforza (1922*)
  - Evolutionary biologist (colleague of Greenberg at Stanford)
  - Human Genome Diversity Project
  - Attempt to develop a biological family tree of modern humankind with the help of genetic analysis
Cavalli-Sforza

• Basic idea:
  – Evolution is based on natural selection: alleles that increase fitness (= ability to survive and replicate) will spread in the population
  – Variation develops via genetic mutations
  – Many mutations do not have any effect on phenotype
  – Therefore no selective pressure
  – Whether or not such a neutral mutation will spread is due to chance => so-called genetic drift
Cavalli-Sforza

• example:
  – Rhesus factor of the blood
  – Has no impact on fitness
  – heritable
  – Connected populations have a specific percentage of Rh-
  – If a population is separated, these values drift apart
  – Difference in percentage of Rh- is thus a crude measure of the relatedness of populations
  – Cavalli-Sforza used several hundreds of such neutral genetic markers
Rekonstruktion der Ausbreitung des Menschen in vorgeschichtlicher Zeit. Ein erster genetischer Stammbaum (grün) wurde derart auf eine Weltkarte projiziert, dass die Endpunkte der Zweige in den heutigen Regionen der einzelnen Populationen liegen. Das Ergebnis passt recht gut zu einer Rekonstruktion nach archäologischen und fossilen Funden (die Zahlen bezeichnen das Auftauchen des anatomisch modernen Menschen in Jahren vor der Gegenwart). Neuere genetische Untersuchungen (rot) lassen vermuten, dass der *Homo sapiens* auf zwei Routen nach Asien gelangte; die Details der Wege beruhen aber auf Spekulation.
Cavalli-Sforza

• Project was partially criticized because it allegedly revived the biological concept of human races

• Criticism is not valid, quite the contrary:
  – Results show that there are no human races in the biological sense (of compete reproductive isolation)
  – Split of human sub-populations occurred a – according to evolutionary standards – very short time ago: about 60,000 years
  – Genetic variation within a population is sometimes larger than between populations
  – Visible features like skin color, hair consistency etc. are determined by small number of genes
Cavalli-Sforza & Greenberg

- Surprisingly good correspondence between Cavalli-Sforza's and Greenberg's classifications of human populations
- Additional argument in favor of Greenberg's results
Cavalli-Sforza & Greenberg

• Minor divergences:
  – Hungarians speak an Uralic language, but are genetically indistinguishable from neighboring populations
  – Sami also speak Uralic language, but are genetically closer related to other Skandinavians and mongoloid siberians
  – Ethiopians speak Afro-Asiatic language, but are genetically closer related to Southern African than to Northern African populations
Using methods from bio-informatics

- problem of reconstructing language trees is similar to problems in evolutionary biology
- bio-informaticians use statistical techniques to induce family relationships between groups of DNA or protein sequences
- these methods are increasingly being applied to linguistic data
Gray & Atkinson (2003) on reconstructing the indo-european tree

- data: **200-word Swadesh list** of 95 indo-european languages (quality is heavily contested, see this Wikipedia entry)

- estimation both of **most likely tree** and most likely time depth of branching nodes

- method is still experimental but has high potential in the future

- main problem: how to distinguish true cognates from borrowings
Sources

- Die Evolution der Sprache, Spektrum der Wissenschaft -- Dossier 1/2000