Santa Fe Institute Seminar
August 12, 2008

Making a Comparative Linguist out of your Computer:

Problems and Achievements

George Starostin

Russian State University for the Humanities

with the programming assistance of
Phil Krylov

Defining comparative linguistics

Systematic comparison of languages to test the hypothesis that they share the same linguistic ancestor ("**genetic relationship**")

Establishing rules responsible for historical transformation of languages ("**regular correspondences**")

Deducing information on the linguistic ancestor of related languages, based on these rules ("**proto-language reconstruction**")

Some typical issues of comparative linguistics

1. How much information is necessary to establish genetic relationship?

Ideal conditions:

— complete phonological and grammatical descriptions of compared languages

— large vocabularies of compared languages to minimize chance coincidences

— knowledge of the general linguistic characteristics of the area to minimize "contact vs. relationship" confusion

Real conditions (in most cases):

— insufficient descriptions, many of them of a dubious nature

— small wordlists of languages, ranging from several dozen to a couple hundred items

— insufficiently explored area with the history and taxonomy of most languages still unknown

The phonological + lexicostatistical criterion of relationship [PLC]

Languages are genetically related, if

a) there is a system of recurrent phonetic correspondences between them

b) this system can be observed on lexical items that belong to the Swadesh 100-wordlist

Example: English vs. German.

bite	beissen
eat	essen
foot	Fuss
water	Wasser

Can be confirmed additionally beyond the 100-wordlist:

goat - Geiss, hate - hassen, nut - Nuss, etc.

Some of the encountered problems of PLC

A. *With a limited amount of data,*
correspondences may be too complex,
sometimes even between closely related languages.

Problem particularly acute in:

— languages with complex phonologies
(North Caucasian, Khoisan etc.)
— morphologically rich languages
(North Caucasian, Niger-Congo etc.)

Example

Some true cognates between Lezghian and Archi
(North Caucasian family, Lezghian branch)

	Lezghian	**Archi**	**Meaning**
1	tssif	dihl^{w}	'cloud'
2	varz	bats	'moon'
3	yab	oy	'ear'
4	vil	lur	'eye'
5	luhu-	bo-	'say'

Items 1-2: complex phonological correspondences.
Items 3-5: similarity obscured by morphological developments.

B. *Languages may be too distantly related*

→ obscure phonetic correspondences
(more time means more phonetic developments)

→ sharp decrease in the number of true cognates
(harder to distinguish from look-alikes or to spot at all)

Example

Hindi vs. English (only 19 out of 100 true cognates)

List includes a few easily recognizable cognates,

English	Hindi
name	nām
no(t)	na
two	do

a few that would be easily spotted by a professional linguist,

English	Hindi
full	pūrā
new [nyu:]	nayā

but most are unrecognizable:

English	Hindi
horn	sīŋg
know	jān-
tongue	jībh-
eye	ākh
one	ek
what	kyā

Normally, such cognates are recognized only if
(a) extra historical information is available
or
(b) additional related languages are found

Another issue:

2. How much depends on the subjective judgement of the historical linguist?

Example: the word "tongue" in Indo-European

Tokharian	käntu
Sanskrit	jihva-
Russian	yazyk
Lithuanian	liežùvis
Gothic	tungō
Latin	lingua
Old Irish	tenge

Correspondences are generally irregular, but most linguists agree the words still share a common origin.

! Problem becomes much harder for more distantly related or less studied languages.

A procedure of automatic computer analysis of comparative wordlists

Primary goals:

— assist the researcher with analyzing data from poorly studied languages

— verify various language relationship hypotheses, including long-range ones

Basic structure of the algorithm

1. Establishes potentially regular correspondences between consonant classes of different languages

2. Suggests cognation between various lexical elements of comparative wordlists

3. Attempts to reconstruct a consonantal "proto-skeleton" for related items

Step I. Establishing potentially regular correspondences between consonant classes of different languages

1. Assigning consonants to various classes:

- based primarily on place of articulation (labials, dentals, velars, etc.)

- disregarding laryngeal features (voiced/voiceless, glottalized, etc.)

Classes preferred to individual consonants so that the algorithm can better operate on small amounts of material.

Example:

Class P	**p, b, ṗ, f, v**
Class T	**t, d, ṭ**
Class K	**k, g, ḳ, x, ɣ**
Class H	**h, ʔ, ZERO**

Same, but with extra differentiation:

Class P	**p, b, ṗ**
Class F	**f, v**
Class T	**t, d, ṭ**
Class K	**k, g, ḳ**
Class X	**x, ɣ**
Class H	**h, ʔ, ZERO**

Examples of representation:
all = HLL; *ashes* [æš-] = HSS; *bark* = PRK; *big* = PK, etc.

2. Calculating frequency of correspondences (pair-wise):

$F_1(P)$ = total number of occurrences of consonantal class P in language 1

$F_2(P)$ = total number of occurrences of consonantal class P in language 2

$F_R(P)$ = total number of occurrences of consonantal class P in the same position for the same lexical item in both languages

The overall frequency is then calculated as

$$2F_R(P) / (F_1(P)+F_2(P))$$

Example (English vs. German)

a) Results for «trivial» correspondences

Pair of classes	F_1 (English)	F_2 (German)	F_R (English-German)	Result
P : P	11	13	5	**0.42**
T : T	43	31	17	**0.46**
N : N	33	47	26	**0.65**

b) Results for non-correspondences

Pair of classes	F_1 (English)	F_2 (German)	F_R (English-German)	Result
P : K	11	25	2	**0.11**
T : M	43	11	1	**0.04**

c) Results for «non-trivial» correspondences:

Pair of classes	F_1 (English)	F_2 (German)	F_R (Engl.-Grm.)	Result
Y : K	21	25	9	**0.39**
T : C	43	8	6	**0.24**
F : P	12	13	2	**0.16**
T : S	43	35	6	**0.15**

List of highest percentages, in decreasing order (Engl.-Germ.)

G : G (ŋ)	**0.80**	T : T (t, d)	**0.46**	T : S	**0.15**
H : H (h, zero)	**0.78**	P : P (p, b)	**0.42**	M : P	**0.14**
R : R (r)	**0.74**	Y : K	**0.39**	Y : S	**0.11**
F : F (f, v)	**0.69**	K : K (k, g)	**0.33**	P : K	**0.11**
S : S (s, z)	**0.68**	T : C	**0.24**	Y : Y	**0.09**
M : M (m)	**0.67**	D (th) : T	**0.21**	H : N	**0.09**
W : W (w)	**0.65**	F : P	**0.16**	Y : H	**0.09**
N : N (n)	**0.65**	P : F	**0.16**	Y : T	**0.08**
L : L (l)	**0.64**	Y : X (ch)	**0.15**	S : K	**0.08**

Step II. Identification of potential cognates

1. Establishing the "threshold of acceptability"

— generally flexible (set as a modifiable parameter)

— calibrated as 0.15 for closely related families (Germanic, Turkic, etc.)

[best results generally yielded with this figure]

— should be increased for distant relatives (Indo-European, Altaic, etc.)

[significant increase of false cognates at 0.15]

Alternatively: distant relatives should not be analyzed this way at all

2. **Building cognate chains**

General transitivity rule:

If A is cognate with B and B is cognate with C, then A is cognate with C

[!!! even if correspondences between A and C are not recognized]

Important corollary:

The accuracy of the results is dependent on the number of languages included.

Example: German vs. English

a) Binary approach

Cognates recognized: 64 out of 100 (all true, no false cognates)
True cognates not recognized: 12 out of 77

b) Multi-language approach
(English and German as 2 out of 12 Germanic languages)

Cognates recognized: 72 out of 100 (all true, no false cognates)
True cognates not recognized: 6 out of 77

How does this happen?

E. g.:

English ***two*** [tū] = TH : German ***zwei*** = CW

In binary comparison, H : W not recognized as a valid correspondence

Adding other Germanic languages:

Danish **to** = TH : English **two** = TH : German **zwei** = CW

Danish H : German W is recognized as a valid correspondence
Danish H : English H is recognized as a valid correspondence

Thus, all three forms are deemed cognate

Some non-identified cognates between English and German even within the multi-language approach:

English	German	Reason
knee [ni:]	Knie	Correspondence "zero : k" not identified
know [nou]	kennen	
mouth	Mund	Correspondence "zero : n" not identified
tooth	Zahn	
woman	Weib	Morphological problems

Additional issue: **Alignment of segments**

General rule:

Items with equal number of consonants follow but one alignment strategy:

for $C_{I-1}C_{I-2}[C_{I-3}] : C_{II-1}C_{II-2}[C_{II-3}]$,

the alignment is:

$C_{I-1} : C_{II-1}$
$C_{I-2} : C_{II-2}$
$C_{I-3} : C_{II-3}$

Items with differing number of consonants follow alternate strategies:

$C_{I-1}C_{I-2}$: $C_{II-1}C_{II-2}C_{II-3}$

Alignment 1: C_{II-3} hypothesized deleted in I or a suffix in II

[German *trocken* - English *dry*]

C_{I-1} : C_{II-1}

C_{I-2} : C_{II-2}

Alignment 2: C_{II-2} hypothesized deleted in I

[German *Nagel* - English *nail*]

C_{I-1} : C_{II-1}

C_{I-2} : C_{II-3}

Alignment 3: C_{II-1} hypothesized deleted in I

[German *Wurzel* - English *root*]

C_{I-1} : C_{II-2}

C_{I-2} : C_{II-3}

Cognate identification stage

Items are identified as cognates if at least one of the strategies leads to a satisfactory result.

Example:

English *root* [RT] : German *Wurzel* [FRC] : Danish *rod* [RT]

Alignment 3: English and Danish roots augmented to HRT

German F : English H *not* recognized
German F : Danish H recognized

By applying the transitivity rule, all three forms are cognate.

Step III. Reconstructing the "proto-skeleton"

Reasons:

[A] Technical: comparison is limited
to 30 languages at a time

[B] Maximizing resemblances
between distantly related languages

1. Most frequent cognate chain(s) selected as representative of the proto-language stage [parameter is flexible]

2. Proto-skeleton reconstructed according to the following rules:

a) identical skeletons are reconstructed the same way;
b) in non-identical skeletons, historic typology of phonetic change is made use of

Example:

German	English	Dutch	Swedish	Proto-skeleton
Zunge	tongue	tong	tunga	*TG
ich	I	ik	jag	*YK
beissen	bite	bijten	bita	*PT

Testing and results

— Has been tested on most major language families of Eurasia

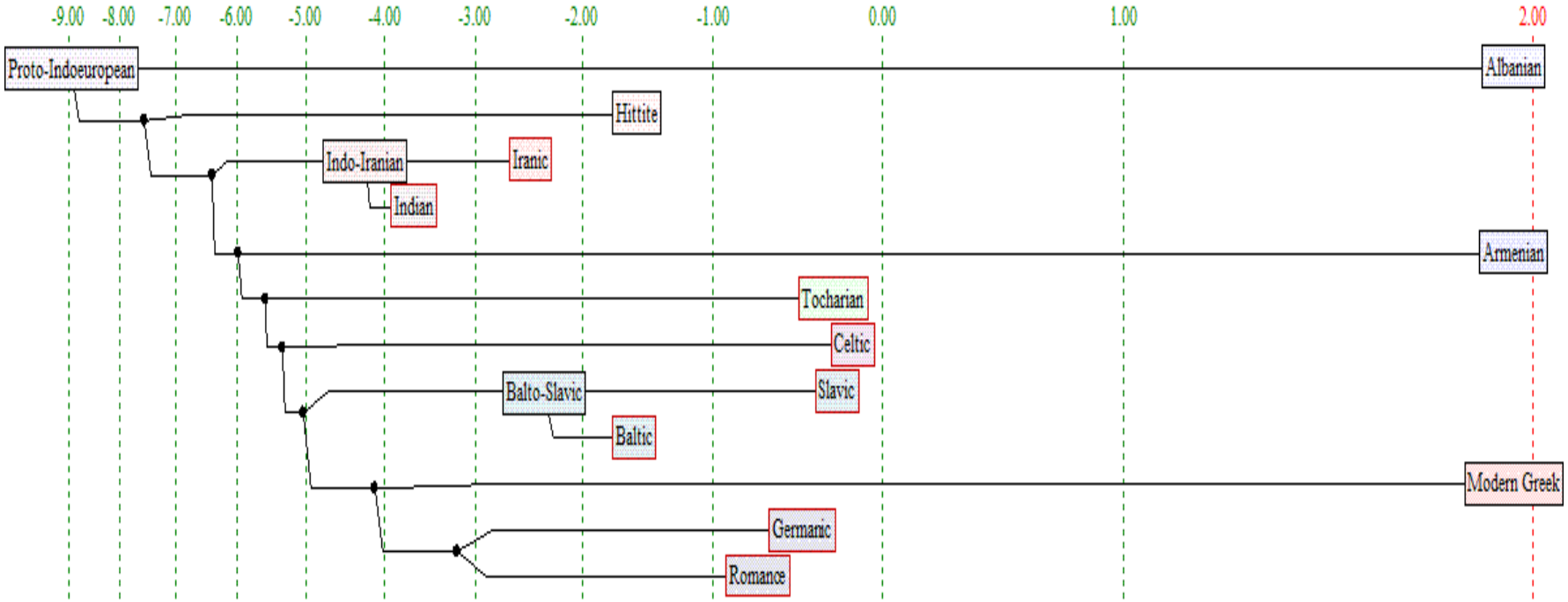
General assessment of results:

"Young" families (2 - 3,000 BP)	Germanic, Slavic, Turkic, Polynesian, etc.	80 - 90% cognates identified correctly; overall classification includes only minor errors; dating of separation almost coincides with regular glottochronological dating
"Medium age" families (4,000 - 6,000 BP)	Indo-European, Uralic, Altaic, North Caucasian, etc.	Well identified on the reconstructed level; 30 - 50% cognates identified correctly; overall classification includes somewhat more errors
"Macro-families" (7,000 BP and higher)	Nostratic, Afro-Asiatic, Dene-Caucasian, etc.	Only partially identified on the reconstructed level; may be due to insufficient data or incompleteness of algorithm

An example of deeper relationship: Modern German vs. Modern Greek [threshold = 0.20]

Cognates recognized (incl. false)			True cognates not recognized		
claw(nail)	Nagel	níxi	*ear*	Ohr	aftí
feather	Feder	fteró	*egg*	Ei	avɣó
fish	Fisch	psári	*eye*	Auge	máti
horn	horn	kérato	*foot*	Fuß	póδi
I	ich	eɣó	*hear*	hören	akúo
knee	Knie	ɣónato	*heart*	Herz	karδía
know	kennen	ɣnorízo	*leaf*	blatt	fílo
name	Name	ónoma	*many*	viel	polí
new	neu	néos	*sand*	Sand	ámos
night	Nacht	níxta	*stand*	stehen	stékome
one	ein	éna	*this*	dieser	aftos
root	wurzel	ríza	*tooth*	Zahn	δódi
seed	Same	spóros	*two*	zwei	δío
star	Stern	astéri, ástro	*what*	was	ti
sun	Sonne	ílios			
thou	du	si			
who	wer	opóios			

Tree diagram of Indo-European (30 modern languages from var. branches)



Future goals

Technical tasks:

— data representation issue
— expanding comparanda beyond the 100-wordlist?..

Substantial tasks:

— further work on the reconstruction part of the algorithm
— separating contacts from cognates