Typology in the service of classification

Johanna Nichols, UC Berkeley

Alternative approaches to language classification

Stanford, July 17-19, 2007

Typology = comparative grammar that deals with types.

Language classification deals with individuals.

How can typology contribute to classification and subgrouping?

Typology = comparative grammar that deals with types.

Language classification deals with individuals.

How can typology contribute to classification and subgrouping?

As heuristic, evaluation measure, and sometimes even firm identification of descent relationships.

But … much purely linguistic work needs to be done first.

Typology = comparative grammar that deals with types.

Language classification deals with individuals.

How can typology contribute to classification and subgrouping?

As heuristic, evaluation measure, and sometimes even firm identification of descent relationships.

But … much purely linguistic work needs to be done first.

- Description
- Data gathering
- Implicational correlations, independence of typological features
- Genealogical stability, diffusibility
- Classification and dating

Genealogical classification by typological characters

Genealogical classification by semi-typological characters

Typology as evaluation metric

Typology and stability of lexical items

What typology can and cannot do

Genealogical classification by typological characters

Genealogical classification by typological characters

The **individual-identifying** statistical threshold:

1/7000 or 0.000143

(since there are about 7000 languages on earth)

plus a conventional level of statistical significance:

0.05	1/350,000	or	0.000 0029	or	3 / 1,000,000
0.01	1/700,000		0.000 0014	or	1 / 1,000,000

Genealogical classification by typological characters

This threshold can be met with shared morphological paradigms:

(1) Germanic suppletive paradigm for 'good' : 'better':

English	*good*	*better*
German	*gut*	*besser*
Swedish	*god*	*bättre*

(2) Gender-number suffixes in Afroasiatic determiners (Greenberg 1960). Analysis (a) treats gender as neutralized in the plural; (b) treats it as syncretized.

	(a) Sg.	(a) Pl.	(b) Sg.	(b) Pl.
Masc.	*-n*	} *-n*	*-n*	*-n*
Fem.	*-t*		*-t*	*-n*

(calculation to follow later)

Genealogical classification by typological characters

Is it possible to define a set of typological characters such that some combinations of their values meet the threshold?

Genealogical classification by typological characters

Is it possible to define a set of typological characters such that some combinations of their values meet the threshold?

Theoretically, yes, but …

- Expected frequencies are defined on the actual frequencies in the world's languages, and this could be a fluke. (Maslova 2000, Nichols 2002)
- Enough of the world's language stocks are isolates or young families that samples are exhaustive rather than representative, so randomization cannot generalize beyond the sample population to anything like "possible human language". (Janssen et al. 2006)
- Sample size (~300 stocks, some geographically non-independent, many underdescribed) is too small for accurate non-randomized significance testing (especially for low-frequency characters, which should be the best diagnostics).

Genealogical classification by typological characters

Can we at least use typological characters as heuristics? as confirmation?

Genealogical classification by typological characters

Can we at least use typological characters as heuristics? as confirmation?

Theoretically, yes, but first we need:

- A good sense of which characters are most and least susceptible to inheritance, spontaneous change (language-internal replacement), diffusion, perseverance in substratum; and how fast they change.
- A polished classification of all languages (stock, subgrouping)
- Reasonably accurate ages for language families (stocks, all subgroups)
- Comprehensive descriptions (grammar, dictionary, corpus) for many languages

Using semi-typological characters to approach the individual-identifying threshold

Using semi-typological characters to approach the individual-identifying threshold

Personal pronoun consonantism (1sg, 2sg):

m-T type: English *me, thee*, Latin acc. *me, te,* Georgian *me, shen,* etc.

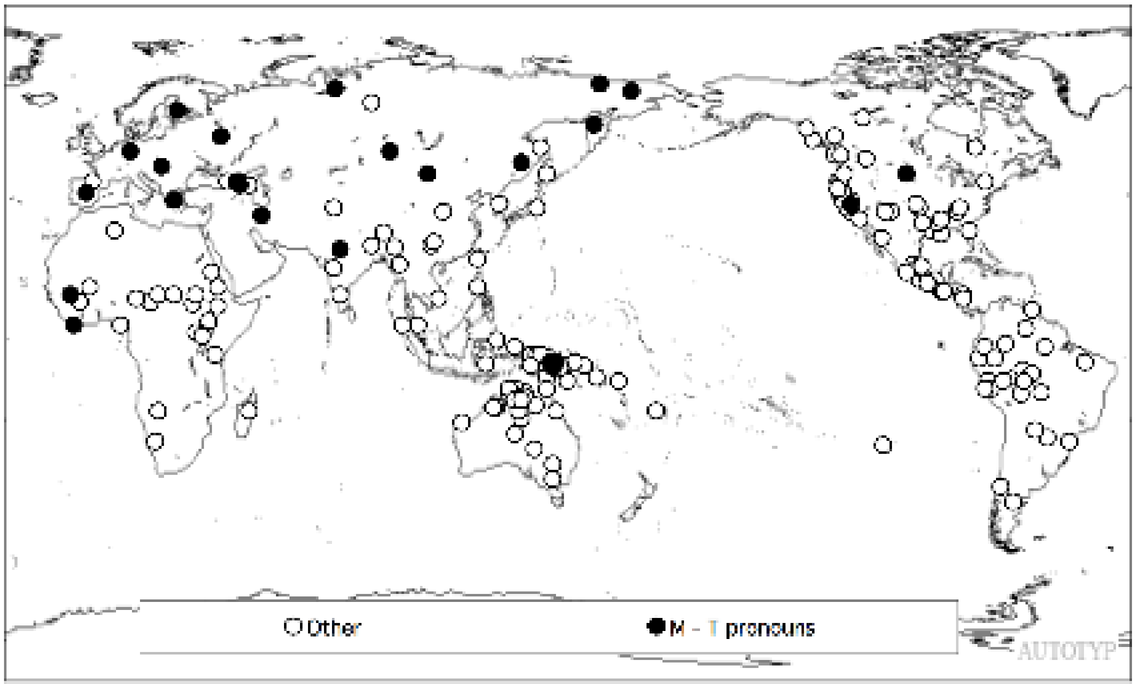
(found in 9 Eurasian stocks)

n-m type: Wintu (Penutian, California) *ni, mi;* Mapudungun (isolate, Chile) poss. *ñi, mi;* etc.

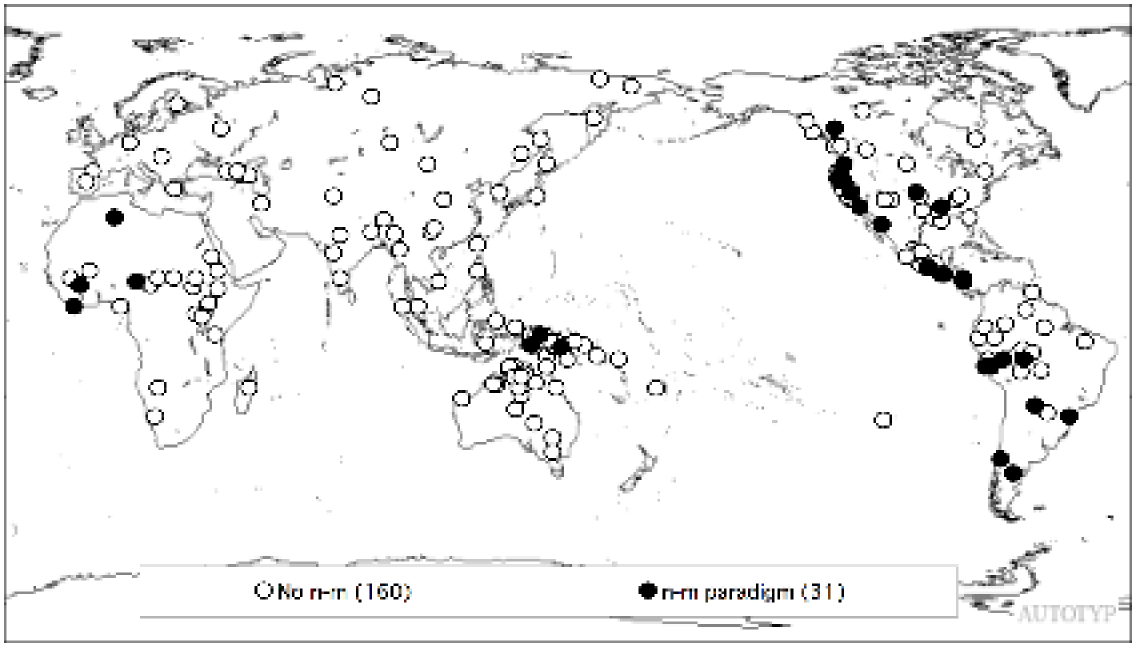
(found in c. 21 American stocks)

(Nichols & Peterson 1996, 2005; Nichols 2001)

Languages with M-T personal pronoun paradigms



Languages with N-M personal pronoun paradigms



Using semi-typological characters to approach the individual-identifying threshold

The geographical distributions show that:

- Both *m-T* and *n-m* systems occur occasionally by chance

- There are two large, high-density clusters

- These clusters must each result from some historical event, connection, relationship, etc.

- We can't determine what that historical situation was: descent? areality? spread of a sound-symbolic canon?

Using semi-typological characters to approach the individual-identifying threshold

Cognate (and putative cognate) roots as **types**

Using semi-typological characters to approach the individual-identifying threshold

Cognate (and putative cognate) roots as types

Two-consonant root :

- C_1 and C_2 (in that order)
- Each C is resemblant (*not* defined by regular correspondences or identity)
- Phonotactics (positioning of vowels, if any) irrelevant

So these represent the same CC root:

- *qof, geb, akpu, xpi* (**similar** consonants)
- plus: *hemi, ogw* (**generic** consonants)

Using semi-typological characters to approach the individual-identifying threshold

Other sources of freedom:

Semantics: same sense; a few senses' leeway; several senses' leeway

Form: strict parse; selective parse

Selective: *kep*, *kedep*, *dekp*, *pek*   (all K-P)

Using semi-typological characters to approach the individual-identifying threshold

Calculation of probability: This is a *search* with several degrees of freedom.

Cumulative probability = $q_1 + q_2 + ... + q_{i+1}$

where $q_j = p\ (1 - q_i)$

p = event probability

q = cumulative probability;

q_i = cumulative probability after the *i*-th trial

Using semi-typological characters to approach the individual-identifying threshold

Example:

Identical (particular) consonant: $p = 0.05$

(Average consonant inventory is about 20.)

Similar consonant: 3 distinctive features' leeway or about 1/7 of consonant inventory: $p = 0.14$

Generic consonant: 5 distinctive features' leeway or about 1/4 of consonant inventory: $p = 0.23$

Similar CC root: $p = 0.02$

Generic CC root: $p = 0.05$

The number of resemblant two-consonant roots required in a binary comparison, with varying degrees of phonological and semantic leeway. Similar calculations for one-consonant roots. (p_2 = probability of two-consonant root; n = number of trials; entries are minimum numbers of words required to reach significance at < 0.05.)

	1 sense:				3 senses:				5 senses:			
	n =	100	200	1000		100	200	1000		100	200	1000
	p_2				p_2				p_2			
Similar	0.02	5	8	28	0.06	10	19	73	0.10	15	28	117
" + select	0.04	7	14	51	0.12	18	33	138	0.18	25	46	201
Generic	0.05	9	16	63	0.14	20	37	159	0.23	30	57	253
" + select	0.09	14	26	106	0.25	32	61	273	0.38	47	88	406
One-consonant roots:												
Generic	0.14	20	37	159	0.37	45	86	396	0.54	64	120	567
" + select	0.27	34	65	294	0.54	64	120	567	0.72	80	155	744

The number of resemblant two-consonant roots required in a binary comparison, with varying degrees of phonological and semantic leeway. Similar calculations for one-consonant roots. (p_2 = probability of two-consonant root; n = number of trials; entries are minimum numbers of words required to reach significance at < 0.05.)

Red = best model of actual long-range comparisons.

	1 sense:				3 senses:				5 senses:			
	n =	100	200	1000		100	200	1000		100	200	1000
	p_2				p_2				p_2			
Similar	0.02	5	8	28	0.06	10	19	73	0.10	15	28	117
" + select	0.04	7	14	51	0.12	18	33	138	0.18	25	46	201
Generic	0.05	9	16	63	0.14	20	37	159	0.23	30	57	253
" + select	0.09	14	26	106	0.25	32	61	273	0.38	47	88	**406**
One-consonant roots:												
Generic	0.14	20	37	159	0.37	45	86	396	0.54	64	120	567
" + select	0.27	34	65	294	0.54	64	120	567	0.72	80	155	744

An example of long-range comparison:

Nikolayev & Starostin's *North Caucasian Etymological Dictionary*

Nakh-Daghestanian (East Caucasian) root: (C)V(R)C

(C1 can be head gender marker)

West Caucasian root: C*(V)

C* = possibly complex

Matching strategy: Multiple selective parse

Match C1 or C2 of EC to any component of C*

If C1 of either language is unmatched it can be considered a gender prefix

Senses: Usually over 5 reported.

3600 reported cognates, 1800 of which have both WC and EC reflexes

No. trials: Wordlist = all available dictionaries for c. 40 languages.

An example of long-range comparison:
Nikolayev & Starostin's *North Caucasian Etymological Dictionary*

Model this search as a binary ND-WC comparison with these parameters:

Consonants: 1 similar (0.14), 1 arbitrary (0.5), total 0.07 for CC root

> (Though in fact the possibility of calling C1 a gender marker makes this de facto not a root consonant, i.e. these are one-consonant roots.)

Selective parse (in addition to the arbitrary C1)

5 senses

Cumulative probability 0.35

Trials: ??? -- Estimate as 7200, twice the number of reported cognates

Successes: **1800** (cognates with WC representatives)

Needed: **2588** (a minimum, as the model above is very conservative)

Another example:

Ruhlen, *PNAS* 1998, Yeniseian - Na-Dene

Putative cognate sets for Proto-Yeniseian and Na-Dene from Ruhlen 1998, classified by phonological structure. All = Na-Dene forms from one or more of Haida, Tlingit, Eyak, Athabaskan. Pr-Ath. = Na-Dene forms from only Proto-Athabaskan.

	All	Pr-Ath. only
2 consonants, strict parse	16	11
2 consonants, selective parse	9	9
1 consonant, strict parse	6	5
1 consonant, selective parse	4	2
0 consonants	1	1
Total	36	28
Total using selective parse	14 (39%)	11 (39%)
Total with 2 consonants	25	20

Another example:
Ruhlen, *PNAS* 1998, Yeniseian - Na-Dene

Parameters of Yeniseian-Athabaskan search:

- 3 senses (most sets contain 2 or 3 different glosses)
- Generic consonants
- 2 consonants (2-cons. sets extracted from the larger corpus)
- Selective parse (used especially for glottal stop, 39% of sets)
- 200-word Proto-Yeniseian wordlist

	Found	Needed
Total sets	28	
Total using selective parse	11 (39%)	
Total with 2 generic consonants	20	37
(needed for selective parse)		61

Another example:
Ruhlen, *PNAS* 1998, Yeniseian - Na-Dene

Parameters of Yeniseian-Athabaskan search:

- 3 senses (most sets contain 2 or 3 different glosses)
- Generic consonants
- 2 consonants (2-cons. sets extracted from the larger corpus)
- Selective parse (used especially for glottal stop, 39% of sets)
- 200-word Proto-Yeniseian wordlist

Additional complicating factor: both compared wordlists are reconstructed protolanguages.

	Found	Needed
Total sets	28	
Total using selective parse	11 (39%)	
Total with 2 generic consonants	20	37
(needed for selective parse)		61

Using semi-typological characters to approach the individual-identifying threshold

Typology as evaluation criterion:

Most long-range comparisons have far fewer proposed cognates than needed.

Most have generous degrees of freedom (phonological, semantic, phonotactic).

Multilateral comparison also has many degrees of freedom in the choice of languages.

Most (all?) offer *only* lexical evidence in support of relatedness.

Using semi-typological characters to approach the individual-identifying threshold

Same evaluation applied to paradigms:

Algic pronominal affixes. I, II = Wiyot allomorph sets.

	Proto-Algonquian	Wiyot I	Wiyot II	Yurok
1st person	* ne-	du(÷)-	d- < *n-	÷ne-
2nd	* ke-	khu(÷)-	kh-	k'e-
3rd	* we-	u(÷)-	w-	÷we- / ÷u-
Indefinite	* me-		b- < *m-	me-

Probability, calculated as 4 identical consonants in a 4-member paradigm:

0.000000024 (2 / 100,000,000)

Same, similar consonants:

0.0000015 (2 / 1,000,000)

Using semi-typological characters to approach the individual-identifying threshold

Germanic *good : better*

English	*good*	*better*
German	*gut*	*besser*
Swedish	*god*	*bättre*

good:	g = 0.05 or 0.14	*bett*-:	b = 0.05 or 0.14
	V = 0.5		V = 0.5
	d = 0.05 or 0.14		t = 0.05 or 0.14
	positive = 0.5		comparative/superlative = 0.5

Overall probability if taken as 4 identical consonants:

0.000 000 39 (4 / 10,000,000)

If taken as 4 similar consonants (p = 0.14 each):

0.000024 (2 / 100,000)

If taken as two similar two-consonant roots:

0.000096 (9.6 / 100,000 or about 1 / 10,000)

Using semi-typological characters to approach the individual-identifying threshold

Gender-number suffixes in Afroasiatic determiners (Greenberg 1960). Analysis (a) treats gender as neutralized in the plural; (b) treats it as syncretized.

(a)	Sg.	Pl.	(b)	Sg.	Pl.
Masc.	-n			-n	-n
		} -n			
Fem.	-t			-t	-n

Probability calculated with specific consonants ($p = 0.05$):

(a)	$p = 0.000\,0045$	(b)	$p = 0.000\,0020$
	(4.5 / 1,000,000)		(2 / 1,000,000)

Probability calculated with similar consonants ($p = 0.14$):

(a)	$p = 0.000099$	(b)	$p = 0.000043$
	(9.9 / 100,000)		(4 / 100,000)

Using semi-typological characters to approach the individual-identifying threshold

Insufficient evidence: *n : m* personal pronoun systems in the Americas

(*n* in 1sg, *m* in 2sg, same paradigmatic positions)

Calculated as 2 identical consonants in a 2-member paradigm:

0.000625 (6 in 10,000)

Same, as 2 identical consonants in particular places in a 6-member paradigm:

0.00007 (7 in 100,000)

Wordlist items in typological perspective

The genealogical stability of words depends on the lexical type of the language.

Wordlist items in typological perspective

The genealogical stability of words depends on the lexical type of the language.

Stance verbs: most stable where the static form is basic; less stable where the dynamic form is basic; least stable where the transitive form is basic.

stand:	static	'stand, be in standing position'
	dynamic	'stand up, get into standing position'
	transitive	'have/make/let stand, stand someone'

'stand' in selected IE branches. (Red: innovations.) (Nichols 2006a, b)

<table>
<tr><th>Branch</th><th>Language</th><th>Static</th><th>Dynamic</th><th>Transitive</th></tr>
<tr><td>Indo-Iranian</td><td>Sanskrit</td><td>stha:-</td><td>stha:-</td><td>stha:-p-aya</td></tr>
<tr><td></td><td>Ossetic</td><td>læwwyn</td><td>styni</td><td>læwwyn kænyn</td></tr>
<tr><td>Slavic</td><td>Proto-Slavic</td><td>sto-j-«e-</td><td>sta(n)-</td><td>stav-i-</td></tr>
<tr><td></td><td>Russian</td><td>stojat'</td><td>vstat' / vstavat'</td><td>stavit' / postavit'</td></tr>
<tr><td></td><td>Polish</td><td>sta|c stoj–e</td><td>(po)wsta|c/(po)wstawa|c</td><td>stawia|c/ postawi|c</td></tr>
<tr><td></td><td>BCS</td><td>stajati stoj—im</td><td>(u)stati/ustajati</td><td>staviti (da stoji)</td></tr>
<tr><td></td><td>Bulgarian</td><td>stoja</td><td>stana;
izpravjam se</td><td>izpravjam;
slo«za 'put'</td></tr>
<tr><td>Italic</td><td>Latin</td><td>sto</td><td>consurgo,
adsurgo, etc.</td><td>pono 'put'
statuo 'put, stand'</td></tr>
<tr><td></td><td>Romanian</td><td>sta</td><td>se scula (în picioare)
se ridica (în picioare)</td><td>scula (în picioare)
ridica 'lift, raise'
pune (pe picioare) 'put'</td></tr>
<tr><td></td><td>Italian</td><td>stare in piedi</td><td>alzarsi (in piedi)</td><td>mettere in piedi
alzare 'lift, raise'</td></tr>
<tr><td></td><td>French</td><td>être debout</td><td>se mettre debout
se lever</td><td>mettre debout
lever 'lift, raise'</td></tr>
<tr><td></td><td>Spanish</td><td>estar de pie</td><td>ponerse de pie
levantarse</td><td>poner de pie
levantar 'lift, raise'</td></tr>
</table>

'stand' in Nakh-Daghestanian languages.

Following Kibrik & Kodzasov 1988, 1990 gender affix is marked with "=".
Blue = archaisms (ancient ND roots).

Branch	*Language*	*Static*	*Dynamic*	*Transitive*
Nakh	Ingush	laatt	ott	otta-=u
Andic	Karata	hercch'e =igh-	hercch'	b=itl-
Lak		=a=c'	=iz	=izan =an
Lezgian	Lezgi	aqqwaz-	qqaragh-	qqaragh-ar-
			(aqqwaz-)	(aqqwaz-ar-)
	Archi	=o=ci	=XXa	ba=XXas a=b=as
	Xinalug	tto:=Xun	tto:=Xun,	
			ttoch	ttoch=Vk

The dynamic form is generally basic, and is innovative in most languages.
Transitive forms are usually derived from dynamic forms.

Conclusions:

Diachronic stability is not a fixed property of particular lexical glosses.

Typology can identify the lexical factors that make particular sets of lexical items more or less stable.

*What strictly typological characters *can* do*

*What strictly typological characters *can* do*

Identify possible and probable sister families.

e.g. Yeniseian and Athabaskan-Eyak-Tlingit

(Vajda 2005, 2006, in press, in prep.)

*What strictly typological characters *can* do*

Identify unsuspected large areas

Continents as areas: Dryer 1989

Transcontinental macroareas:

Circum-Pacific

Pacific Rim

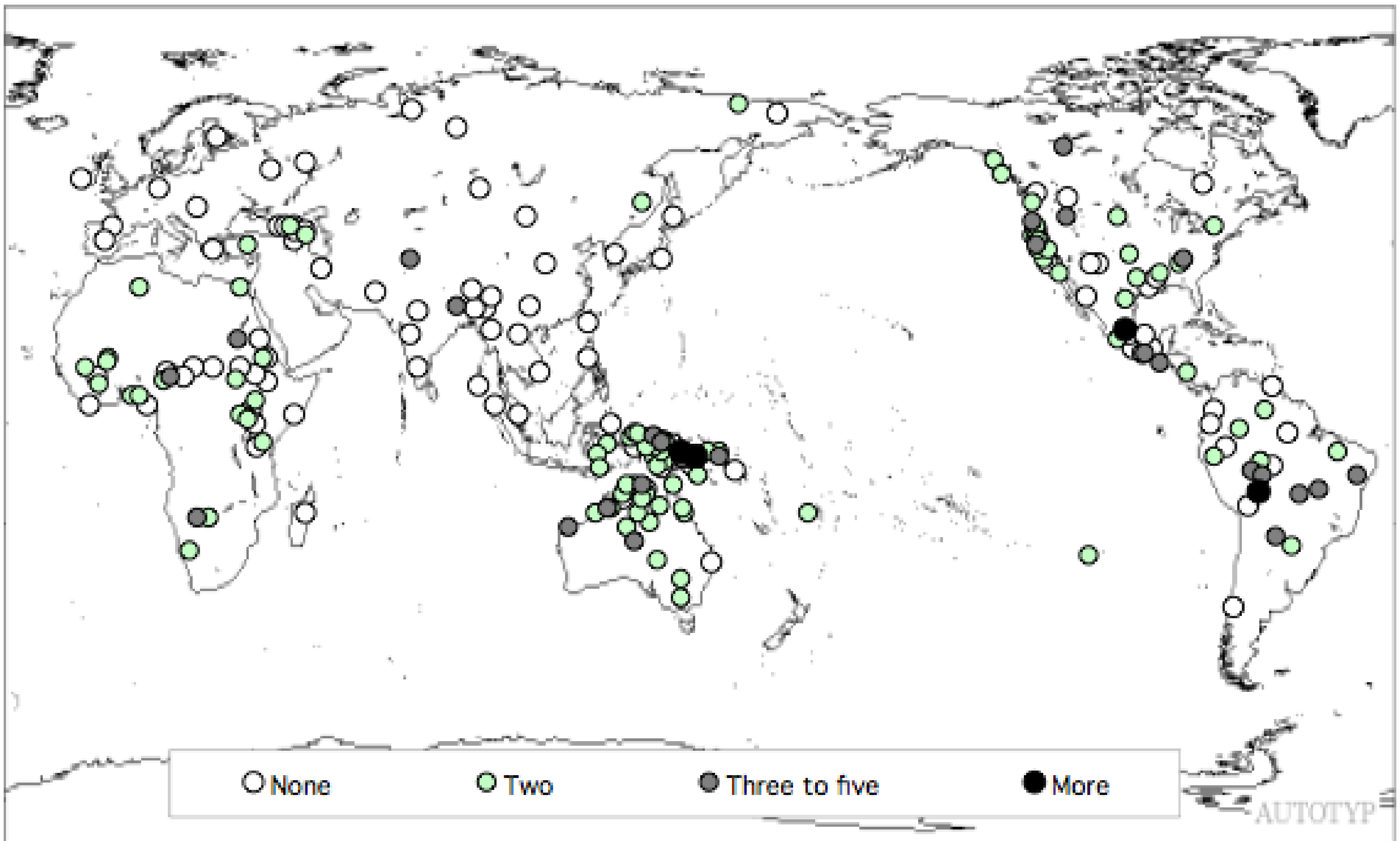
Silk Road

Caucasus-Himalayas

(Nichols 1994, 1997, Nichols & Peterson 1996, Bickel & Nichols 2005, 2006, in prep.)

AUTOTYP: http://www.uni-leipzig.de/~autotyp/)

Number of overt possessive classes



*What strictly typological characters *can* do*

Remove supposed areas

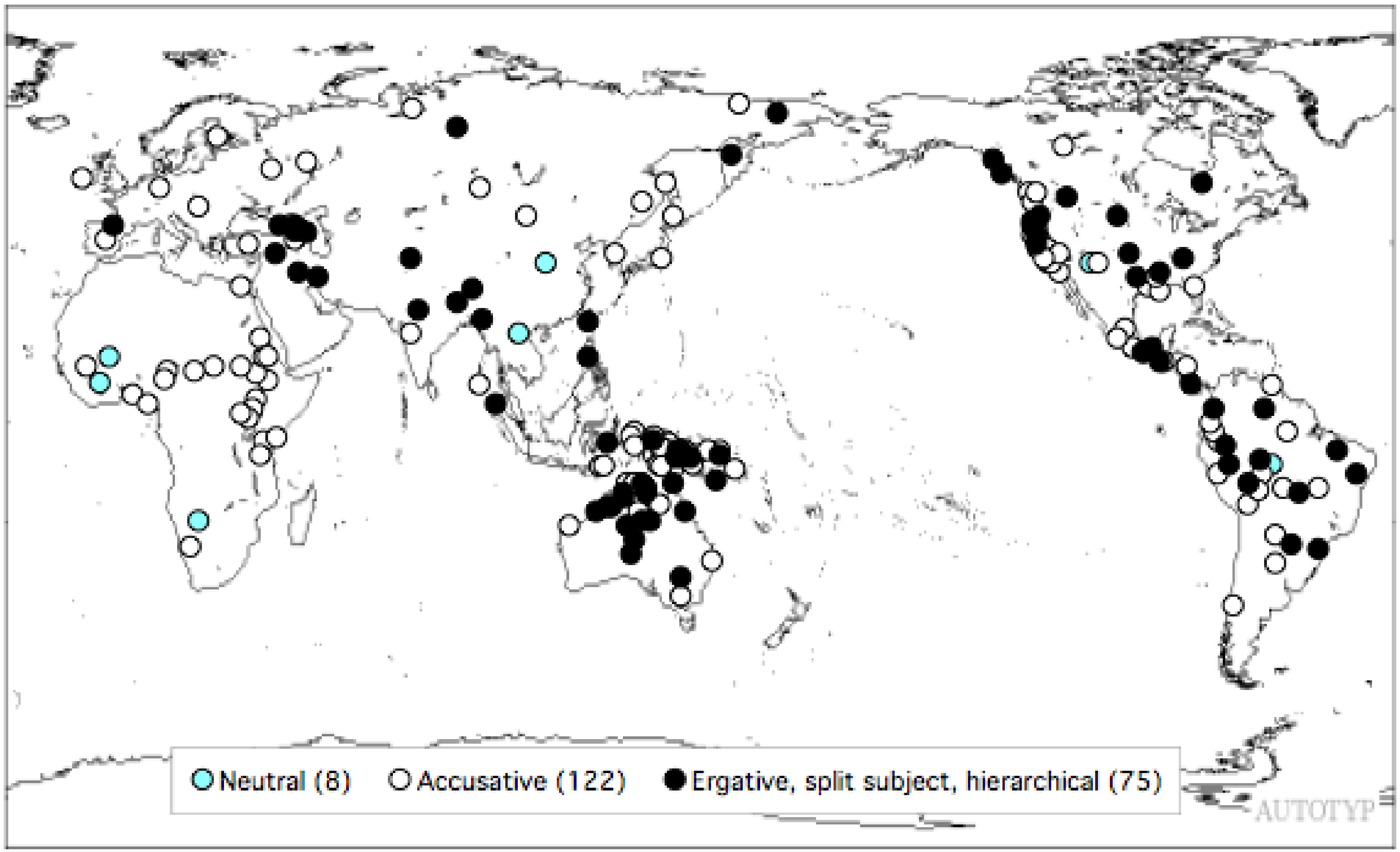
e.g. Eurasia (chiefly northern): not really an area; just skewing within families which have spread widely for economic reasons.

A standard genealogical sample overrepresents these families, all of which are internally quite uniform.

Bickel in press, Bickel & Nichols 2005, 2006

AUTOTYP: http://www.uni-leipzig.de/~autotyp/

Dominant alignment (N = 205)



*What strictly typological characters *can* do*

Point to probable earlier areal connections

- Ket and southern Eurasia (Vajda n.d., Nichols in press)
- Indo-European and northern Eurasia (Nichols in press)
- Munda and Himalayas (Bickel 2005)

*What strictly typological characters *can* do*

These macroareal connections are older than the oldest known stocks, but typological comparison cannot tell us whether their genesis was genealogical or areal.

Conclusions

Conclusions

Standard comparative-historical method identifies and describes particular individuals (language families). Excellent resolution up to the stock level.

Conclusions

Standard comparative-historical method identifies and describes particular individuals (language families). Excellent resolution up to the stock level.

Typology can go much farther back in time, but for purposes of discriminating genealogical from other relatedness it has weak resolution at all time depths.

Conclusions

Standard comparative-historical method identifies and describes particular individuals (language families). Excellent resolution up to the stock level.

Typology can go much farther back in time, but for purposes of discriminating genealogical from other relatedness it has weak resolution at all time depths.

The weak resolution is not inherent; it is due to our primitive understanding of different kinds of diachronic stability, interdependence of characters, rates of change, etc. and our incomplete classification and dating of families.

There is much linguistic work to do before we will have a good set of comparanda.

Conclusions

Standard comparative-historical method identifies and describes particular individuals (language families). Excellent resolution up to the stock level.

Typology can go much farther back in time, but for purposes of discriminating genealogical from other relatedness it has weak resolution at all time depths.

The weak resolution is not inherent; it is due to our primitive understanding of different kinds of diachronic stability, interdependence of characters, rates of change, etc. and our incomplete classification and dating of families.

There is much linguistic work to do before we will have a good set of comparanda.

We can't hope to push the limits of the comparative method back very far.

At all times, whatever the state of knowledge, *the oldest detectable historical connections will always be ambiguous:* genealogical? areal? both? other?

References

Bickel, Balthasar. In press. A refined sampling procedure for genealogical control. *Sprachtypologie und Universalienforschung.*

Bickel, Balthasar, and Nichols, Johanna. 2003. Typological enclaves. Paper presented at *5th biannual conference, Association for Linguistic Typology*, Cagliari, Sardinia.

Bickel, Balthasar, and Nichols, Johanna. 2005. Inclusive/exclusive as person vs. number categories worldwide. In *Clusivity*, ed. Elena Filimonova, 47-70. Amsterdam/Philadelphia: Benjamins.

Bickel, Balthasar, and Nichols, Johanna. 2006. Oceania, the Pacific Rim, and the theory of linguistic areas. *BLS* 32.

Dryer, Matthew. 1989. Large linguistic areas and language sampling. *Studies in Language* 13:257-292.

Greenberg, Joseph H. 1960. An Afro-Asiatic pattern of gender and number agreement. *Journal of the American Oriental Society* 80:317-321.

Janssen, Dirk P., Bickel, Balthasar, and Zuniga, Fernando. 2006. Randomization tests in language typology. *Linguistic Typology* 10:419-440.

Kibrik, A. E., and Kodzasov, S. V. 1988. *Sopostavitel'noe izuchenie dagestanskix jazykov: Glagol*. Moscow: Moscow University.

Kibrik, A. E., and Kodzasov, S. V. 1990. *Sopostavitel'noe izuchenie dagestanskix jazykov: Imja. Fonetika.* Moscow: Moscow University.

Maslova, Elena. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4:307-333.

Nichols, Johanna. 1997. Modeling ancient population structures and movement in linguistics. *Annual Review of Anthropology* 26:359-384.

Nichols, Johanna. 1997. Sprung from two common sources: Sahul as a linguistic area. In *Archaeology and Linguistics: Aboriginal Australia in Global perspective*, eds. Patrick McConvell and Nicholas Evans, 135-168. Melbourne: Oxford University Press.

Nichols, Johanna. 2001. Why "me" and "thee"? In *Historical Linguistics 1999*, ed. Laurel J. Brinton, 253-276. Amsterdam-Philadelphia: Benjamins.

Nichols, Johanna. 2002. Monogenesis or polygenesis? Typological perspective on language origins. Plenary lecture presented at LSA Annual Meeting, San Francisco.

Nichols, Johanna. 2005. The origin of the Chechen and Ingush: A study in alpine linguistic and ethnic geography [2004]. *Anthropological Linguistics* 46:129-155.

Nichols, Johanna. 2005. Quasi-cognates and lexical type shifts: Rigorous distance measures for long-range comparison. In *Phylogenetic Methods and the Prehistory of Languages*, eds. James Clackson, Peter Forster and Colin Renfrew, 57-65. Cambridge: McDonald Institute for Archaeological Research.

Nichols, Johanna. 2006. Stance verbs and the sociolinguistics of the Slavic expansion. Paper presented at *Slavic Linguistics Society inaugural meeting*, Bloomington.

Nichols, Johanna. 2007. A typological geography for Indo-European.

Nichols, Johanna, and Peterson, David A. 1996. The Amerind personal pronouns. *Language* 72:336-371.

Nichols, Johanna, and Peterson, David A. 2005. Personal pronouns: M-T and N-M patterns. In *World Atlas of Language Structures*, eds. Martin Haspelmath, Matthew Dryer, Bernard Comrie and David Gil, 544-551. Oxford: Oxford University Press.

Nikolayev, S. L., and Starostin, S. A. 1994. *A North Caucasian Etymological Dictionary*. Moscow: Asterisk.

Vajda, Edward. 2003. Ket verb structure in typological perspective. *Language Typology and Universals* 56:55-92.

Vajda, Edward. 2005; 2006; in press