Interrelationships between measures of phonological complexity:

Some analysis of the relationship between syllable structures, segment inventories and tone contrasts

Ian Maddieson

University of California, Berkeley

Workshop on Phonology and Complexity, DDL, Lyon, July 4-6, 2005

also ILPGA, Paris, March 11, 2005, ZAS, Berlin, Feb. 11, 2005 and LSA Annual Meeting, Jan. 8, 2005

Complexity

Today's topic: discussion of ‘complexity’ of several properties of phonological systems

globally, are these positively or negatively correlated?

— focus on syllable canon; how does complexity of syllable canon relate to size of consonant inventory, size of vowel quality inventory, complexity of tone system, and how do these other properties relate to each other?

Complexity

Often assumed that greater complexity in one subsystem will be compensated by greater simplicity in another, either because that is the way historical processes work ...

..... And Change said, "Let the consonants guarding the vowel to the left and the right contribute some of their phonetic features to the vowel in the name of selfless intersegmental love, even if the consonants thereby be themselves diminished and lose some of their own substance. ...." (Matisoff 1973)

Complexity

... or because “all languages are equally complex” (this argument may be doctrinaire, or related to processing or memory limitations, etc)

..... all known languages are at a similar level of complexity and detail, ...” (Akmajian et al 1997)

Many possible conceptions of what constitutes complexity (Kusters & Muysken 2001) — here simple notion that increasing number of distinctions equates with increasing complexity

Lack of data richness and reliability often restricts possible comparanda over large number of languages

Wide range of language variation also demands ‘data reduction’ to make analysis feasible

Today, a report on ongoing expansion of what began life as UPSID (Maddieson 1984) and interim analyses of the data

Language Sample

- Union of:
  - ”Enlarged UPSID" language sample (451 languages, aimed at genetic balance) and WALS 200 language sample (geo-political coverage, grammar availability, presence in other samples)
  - plus "Syllables" 30-language database and "Phonetic studies of endangered languages" targets
  - plus other additions (“UPSID Plus”), in progress towards 1000 language target total
- Currently — 614 languages (not more, due to overlap in sample coverage), at various levels of completeness

Language Sample

- Original UPSID database contained only segmental data
- Current sample expanded in content to include information on syllable structure, tone system, and stress, and to represent directly many derivative properties (e.g. presence/absence of voicing contrast in obstruents)
- Merging and expansion of samples relaxes a restriction on inclusion of closely related languages in original design, and exacerbates problems of interpreting results — since representativeness of sample overall, and independence of individual sample languages are both degraded)

An earlier study of syllable patterns in the lexicons of 30 languages (Maddieson 1992)* showed the utility of a basic three-way grouping of syllable types both within the lexicon of individual languages and when comparing across languages (*"The structure of segment sequences" Proceedings of the 1992 ICSLP (Banff, Alberta), Addendum. 1-4)

Simple: no coda, maximum onset one consonant — CV, V (e.g. Yoruba, Maori)

Moderately complex: maximum coda one consonant and/or maximum onset obstruent + 'glide' or 'liquid' — VC, CVC, CGV, CLV, CGVC, etc (e.g. Mandarin, Nahuatl)

Complex: more than one consonant in coda and/or more complex onsets than above — VCC, CCV, CCCVCCC, etc (e.g. Georgian, German)

For this presentation languages are classified into types based on the most elaborate of these syllable types they permit (in established lexical items, disregarding recent 'international' loans — generally no account taken of the relative frequency of the types)

Values for ***syllable complexity*** established for 507 languages:

Simple: 62 languages (12.2%)

Moderately complex: 288 languages (56.8%)

Complex: 157 languages (31.0%)

The syllable complexity classes are compared with:

Consonant inventory size: **Numeric**: Total number of distinctive consonants recognized for the language

Vowel quality inventory size: **Numeric**: Number of distinct vowel types contrasting on major parameters ***(independent of length, nasalization, voice quality, etc)***

Tone system complexity: **Categorical**: Languages sorted into three categories: 1 = no tones, 2 = simple tone system (basic 2-level system), 3 = complex tone system (three or more basic contrasts, and/or contour tones)

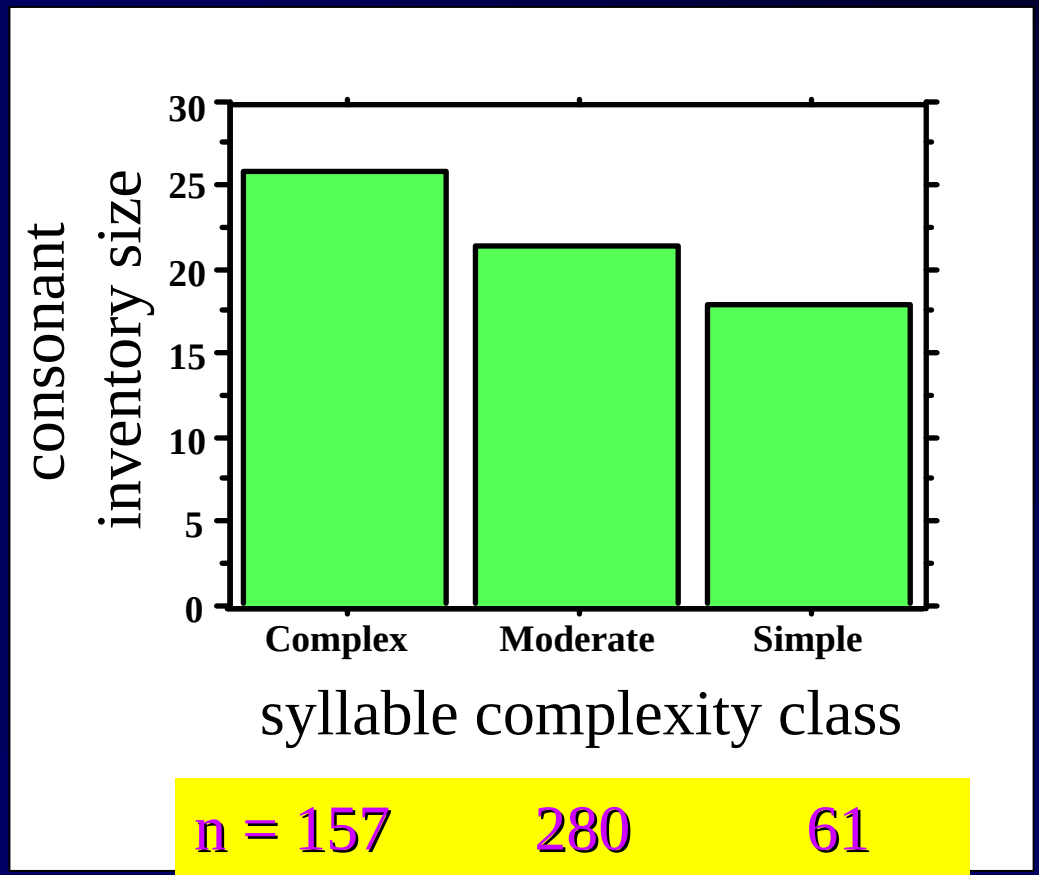
These relationships between the factors will be first discussed using all the languages in the sample for which the relevant data are currently entered in the database

Subsequently some of the issues relating to validating the global results will be examined

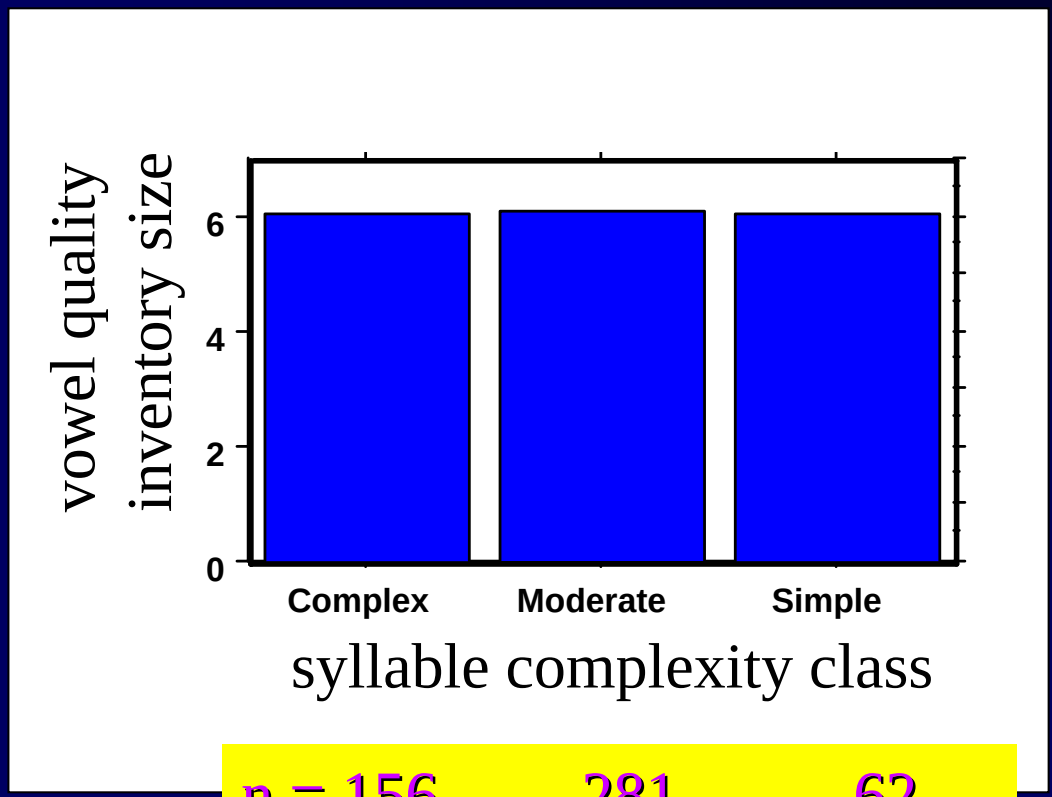
First comparison examines mean size of consonant inventory according to the syllable-structure categories

Complexity of ***syllable structure*** and mean size of ***consonant inventory*** **correlate positively** with each other, contrary to compensation hypothesis. Sample size 498. Difference between classes highly significant

3 outliers with 70 or more consonants omitted from calculation



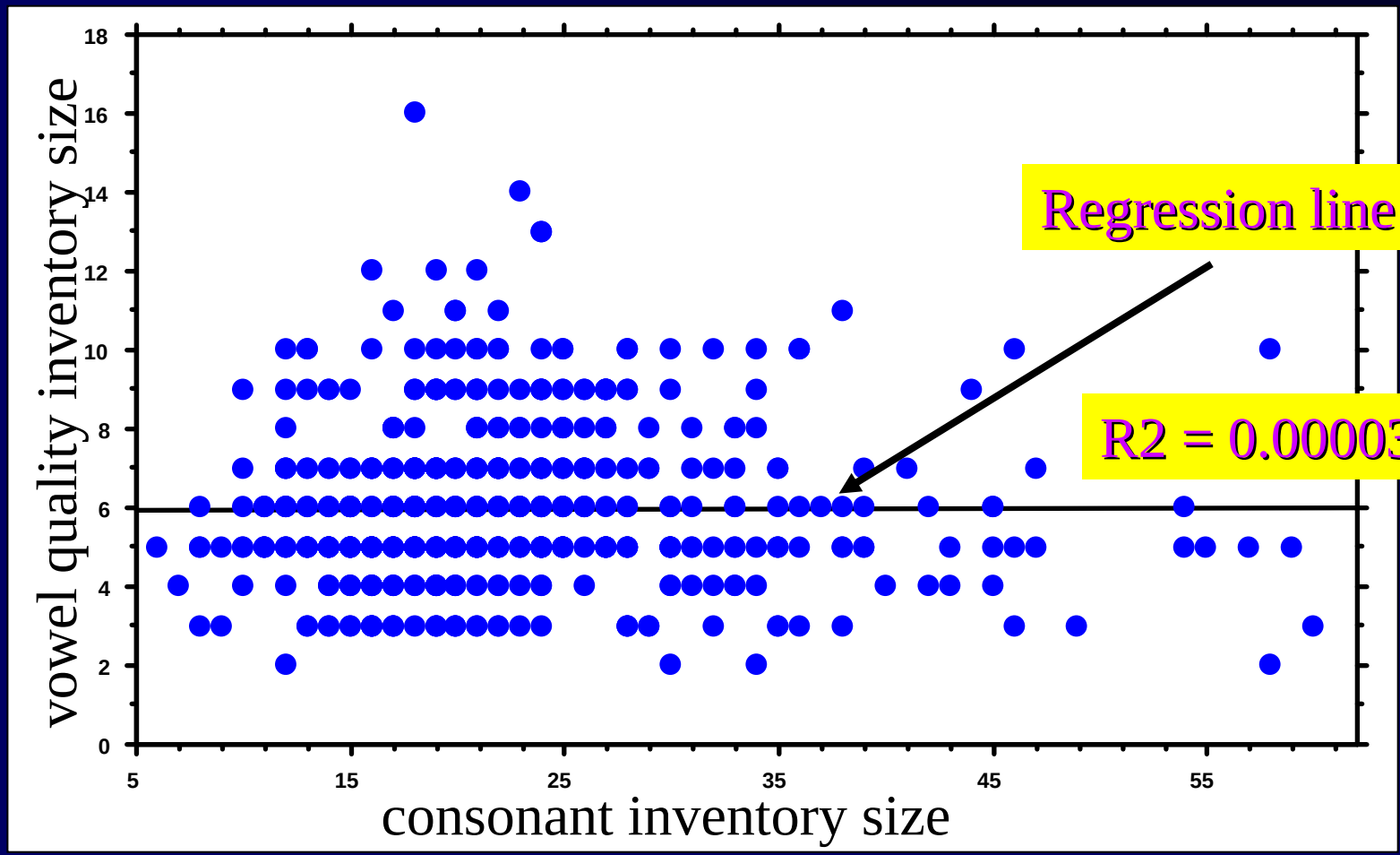
Complexity of ***syllable structure*** and mean size of ***vowel quality inventory*** **do not correlate** with each other either way. Sample size 499. No significant differences



n =	156	281	62
$\bar{x}$ =	6.00	6.07	6.03

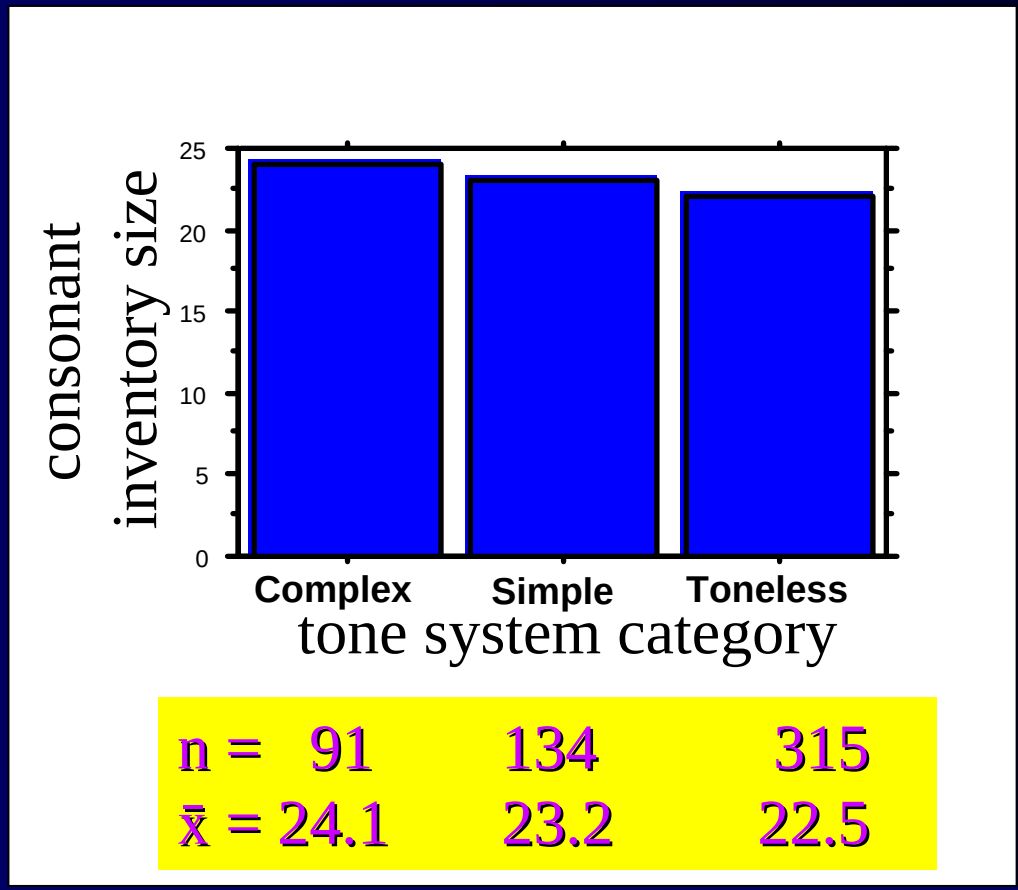
Size of ***consonant inventory*** and size of ***vowel quality inventory*** (numerical measures) absolutely **do not correlate** with each other. Sample size 524

3 outliers with 70 or more consonants omitted from calculation

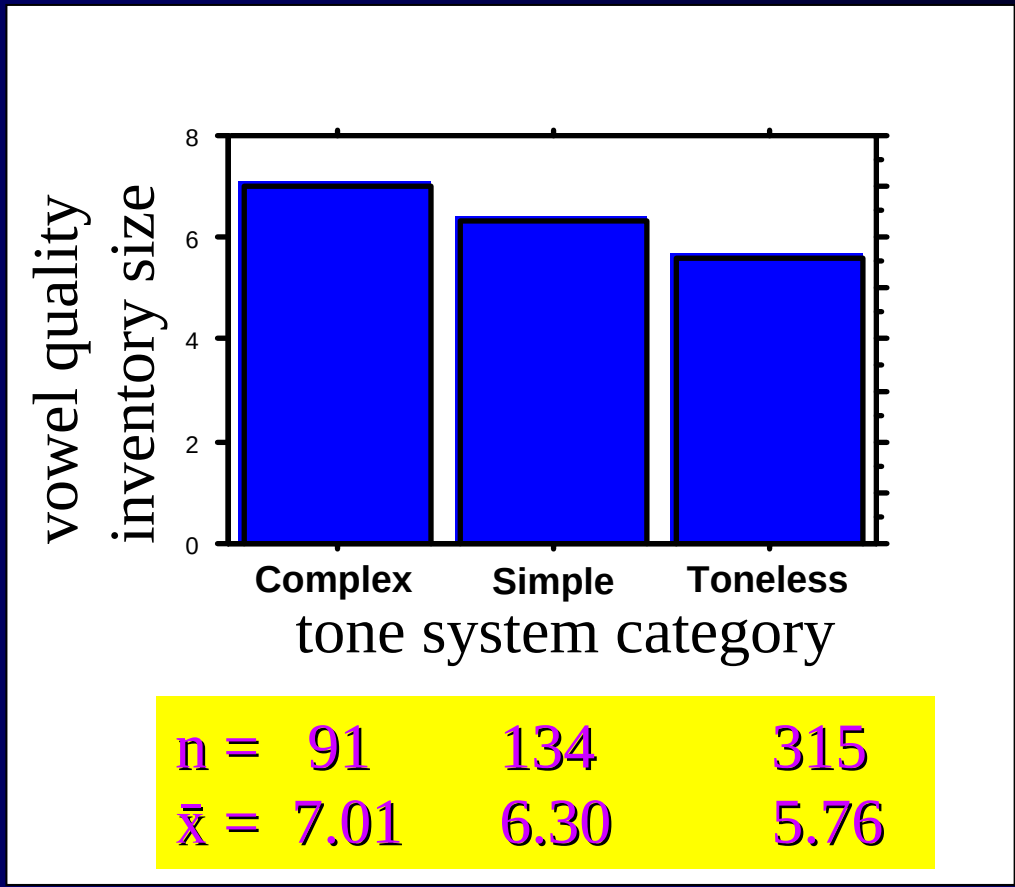


Size of ***consonant inventory*** and complexity of ***tone system*** **positively correlate**. Difference not significant, but clearly not 'compensatory' relationship. Sample size 540

3 outliers with 70 or more consonants omitted from calculation

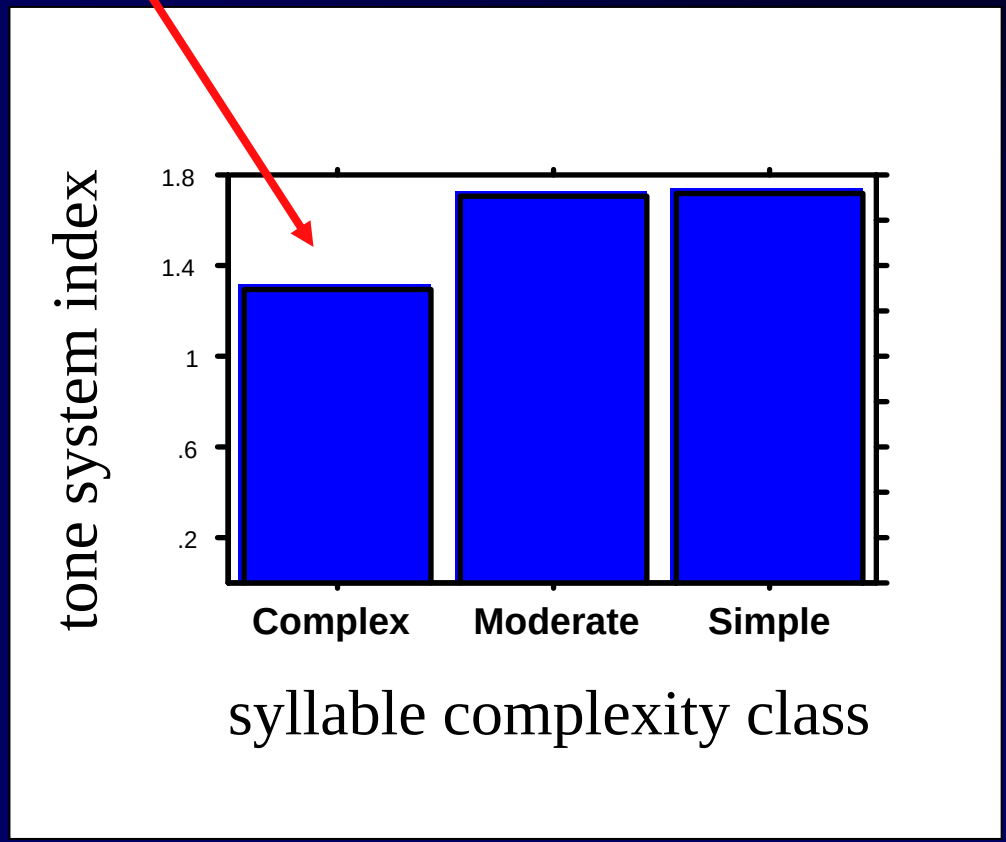


Size of ***vowel quality*** inventory and complexity of ***tone system*** **positively correlate** with each other. Difference highly significant. Sample size 540



Tone system complexity shows **some negative correlation** with ***syllable structure*** complexity; significant difference between complex syllable class and others. Sample size 485

tone system index calculated as 3 = complex, 2 = simple, 1 = none



Summary (1)

Increasing syllabic complexity is positively associated with
— increase of consonant inventory

Increasing elaboration of tone system is positively associated with
— size of consonant inventory
— size of vowel quantity inventory

Summary (2)

Increasing syllabic complexity is unrelated to
— size of vowel quality inventory

Increasing elaboration of consonant inventory is unrelated to size of vowel quantity inventory

Increasing syllabic complexity is negatively related to complexity of tone system

Summary (3)

Of the six relationships studied here, only one indicates any ‘compensation’; three show higher complexity associated with higher complexity, two show no overall compensation

Overall impression is that languages certainly do not systematically 'balance' complexity of one part of phonological system by simplifying elsewhere

Validation

These results are over the surveyed languages as a whole — can they be taken as ***general*** or do they reflect strong but ***local*** tendencies, patterns that are particularly strong in particular areas or language families, but not general

Two methods have been suggested to test the generality of results in large language samples

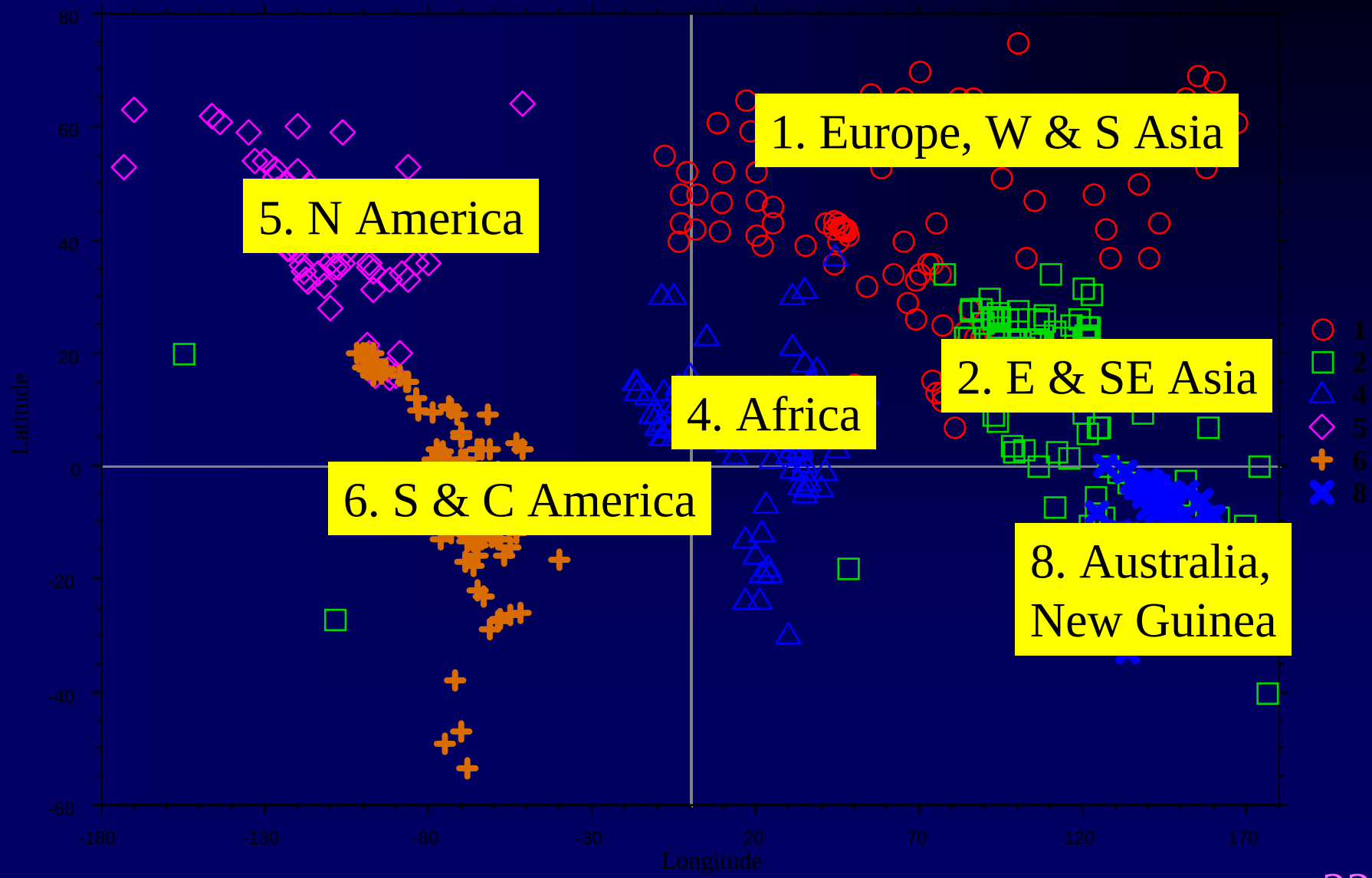
One: divide the sample on areal/genetic lines and look for the repetition of a pattern across the divisions (cf Dryer 1999). Two: create small (sub)samples of genetically/areally independent languages and look for appearance of the effect in these (sub)samples (Stephens & Justeson 19xx)

Validation

Sample divided into 6 ***jointly*** areal+genetic groupings; based on continental landmasses and major families

1. Europe, West & South Asia: Indo-European, Dravidian, Altaic, Caucasian groups, etc — 93
2. East & South-East Asia: Sino-Tibetan, Austro-Asiatic, Tai-Kadai, Austronesian, etc — 107
4. Africa: Niger-Congo, Nilo-Saharan, Afro-Asiatic, "Khoisan" — 144
5. North America: Na-Dene, Eskimo-Aleut, Uto-Aztecan, Algonquian, "Hokan", etc — 75
6. South & Central America: Cariban, Arawakan, Tupian, Mayan, Oto-Manguean, etc — 106
8. Australia & New Guinea:Australian, ‘Papuan’ languages — 89

'Ghost' map of the languages classified by major areal/genetic groups



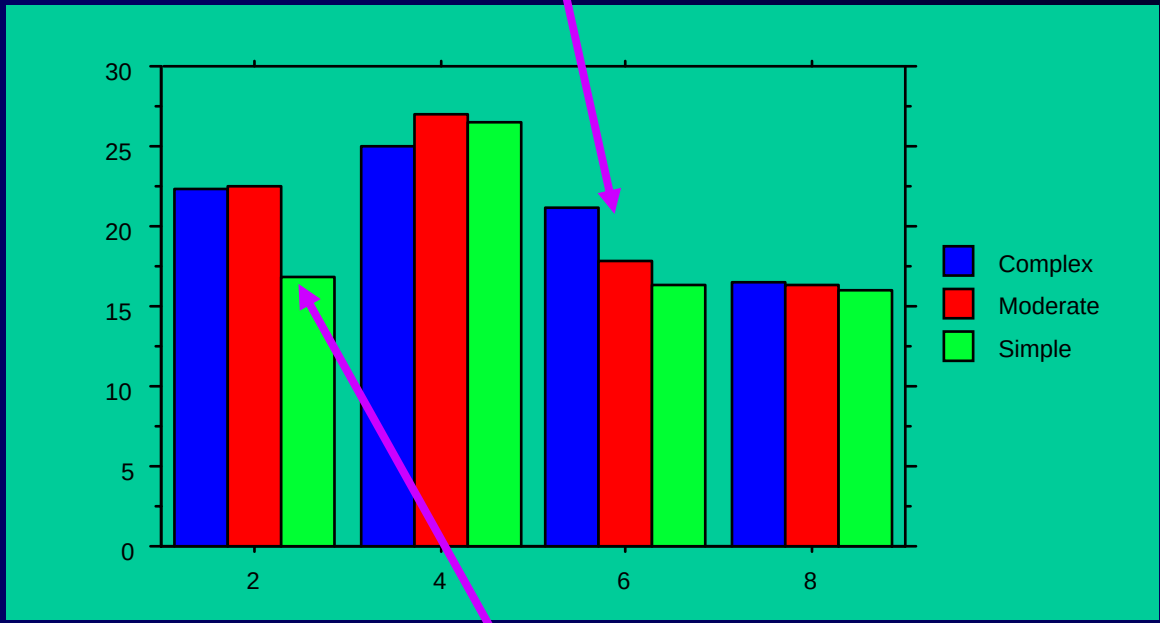
To verify the overall pattern of ***syllable-structure***/***consonant inventory*** independently for each group, values are required in a matrix of 18 cells (6 groups x 3 syllable structures). But 'simple' syllable-structure languages are unequally distributed between groups, and occur neither in Group 1 (Europe, W & S Asia), nor in Group 5 (N America).

# of languages in cell

	1	***2***	***4***	***5***	***6***	***8***
Complex	55	18	24	30	15	15
Moderate	19	66	82	25	42	48
Simple	0	10	18	0	23	11

It is thus not possible to test for the overall pattern in ***all*** groups

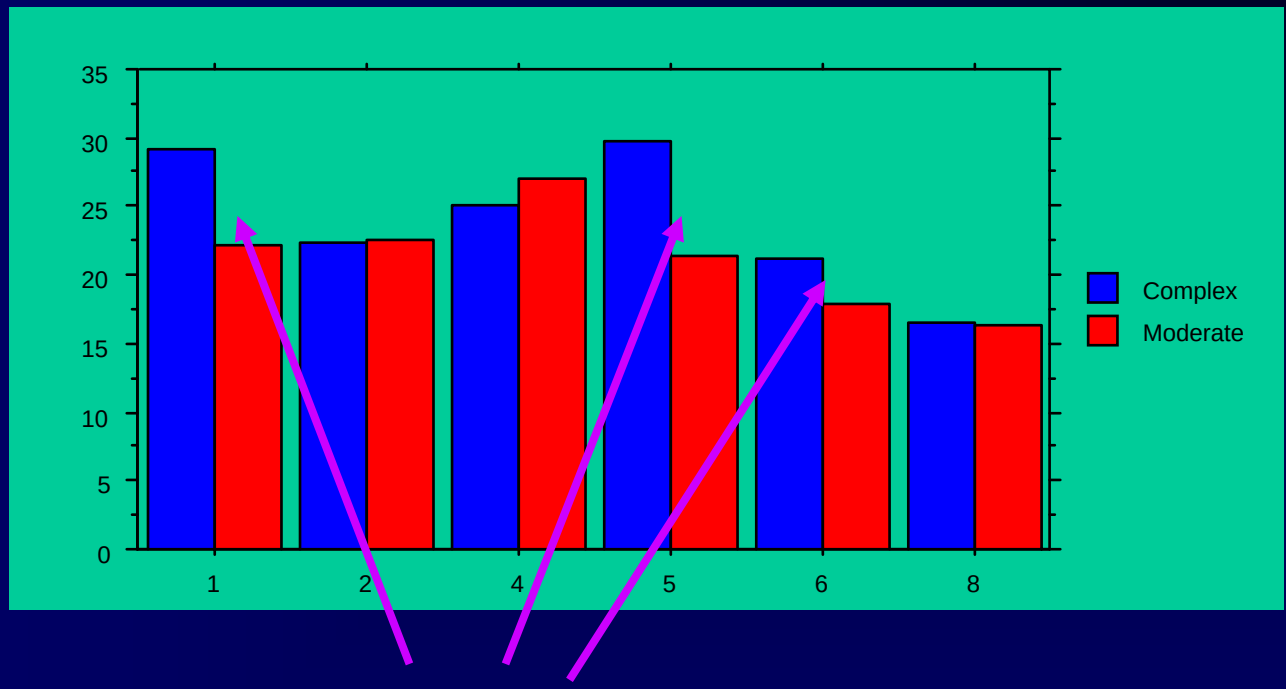
In the four remaining groups only one (6: S & C America) has a monotonic decrease in average consonant inventory size as maximum syllable complexity increases



Group 2 (E & SE Asia) has lower mean for 'simple' than others

Remaining two groups are inconclusive, but do not show persuasive evidence of compensation

Difference between languages with 'moderate' and 'complex' syllable structure can be examined in all groups



'Complex' languages in 3 groups (1, 5, 6) show more consonants than 'moderate' ones

The other three groups are inconclusive, but do not show persuasive evidence for compensation

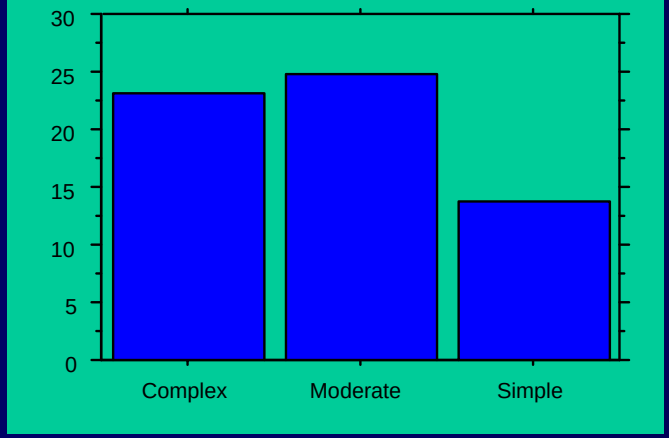
Validation

Attempt to validate the positive correlation of syllable complexity and consonant inventory size by analyzing total sample by groups provides only weak support

Second method: examine small samples drawn from total set — target around 30 to avoid lack of sample independence

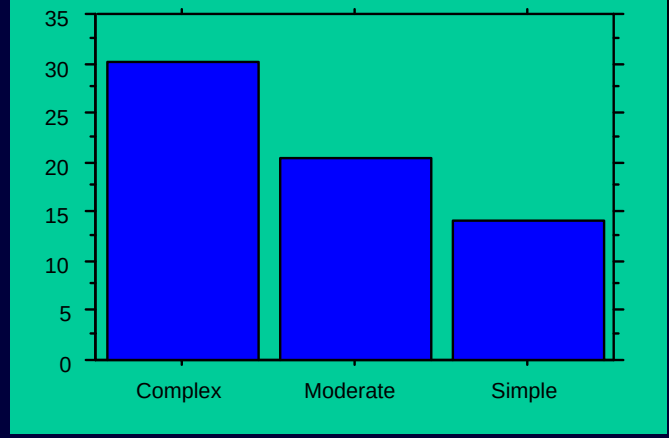
Subsamples can be randomly drawn — probability that language pairs undesirably close genetically or geographically are drawn is low if whole sample is broad-based

Samples can be drawn by quota principles

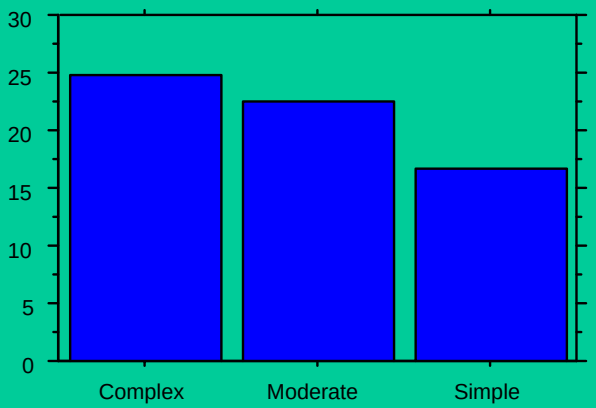


23.1 24.8 13.8

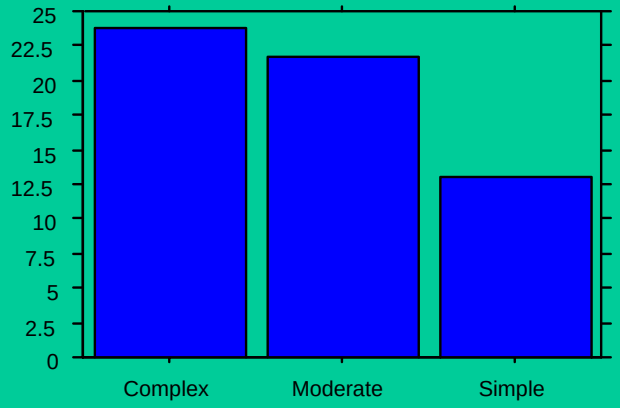
Five random selections of 30 languages, unconstrained



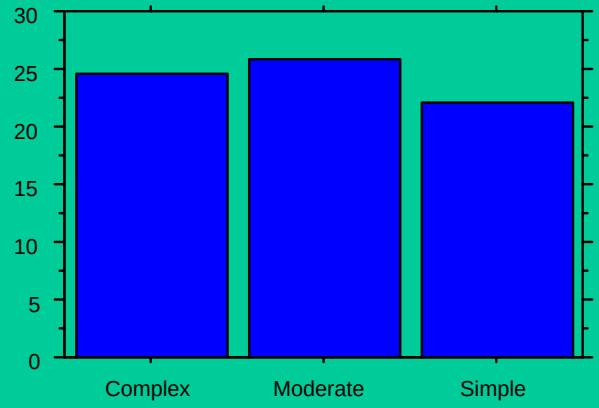
30.1 20.4 14.0



24.8 22.5 16.6



23.8 21.6 13.0



24.7 25.9 22.0

Validation

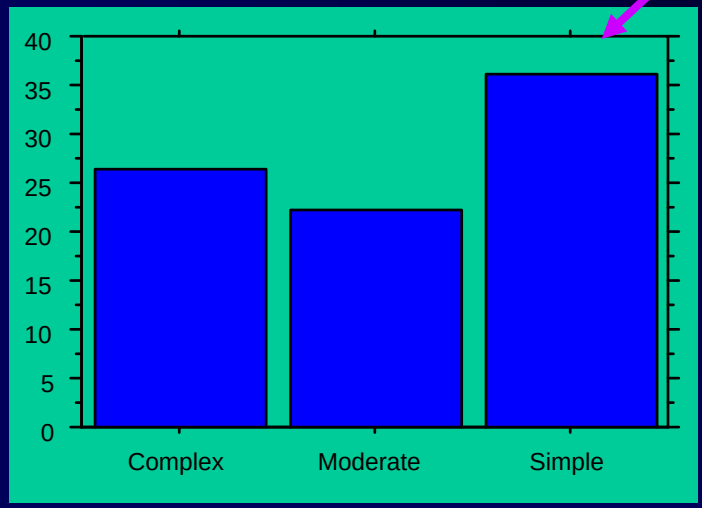
Majority of random samples confirm positive correlation of syllable complexity and consonant inventory size, but single random samples give inconsistent results

Taking many random subsamples gives a 'majority vote', but ultimately blurs the distinction with examining the entire sample

Subsamples can instead be selected using a quota principle, which constrains the selection of languages from separate areas/families

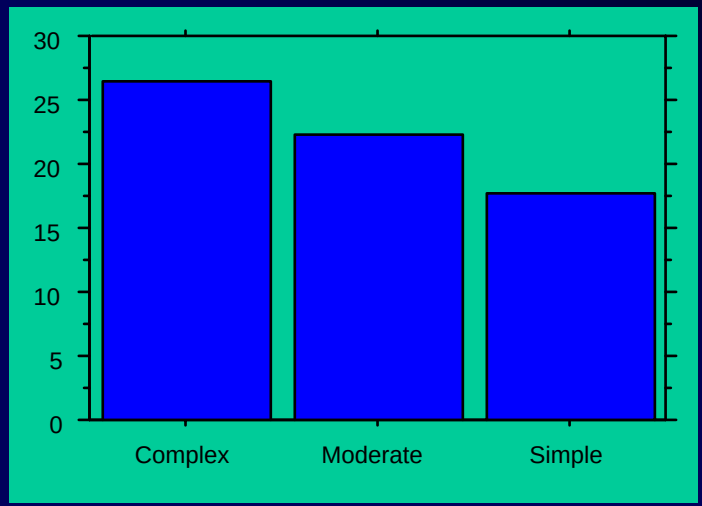
Quota sample:
Guided selection of 32 languages, based on list used by Shosted (2004) using language areas from Nichols (2003) with 2 substitutions for languages with incomplete data

Simple syllabic-structure languages have markedly larger consonant inventory!



26.5 22.2 36.3

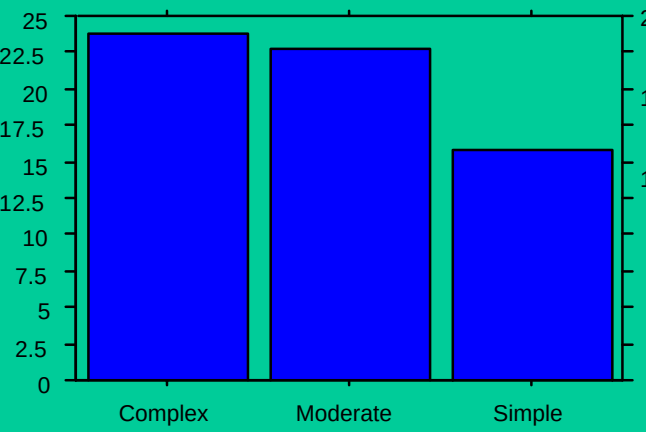
... But, this result only occurs because Ju|hoan — one of the marked outliers in consonant inventory size — is included in the sampled set of only 4 ‘simple’ languages

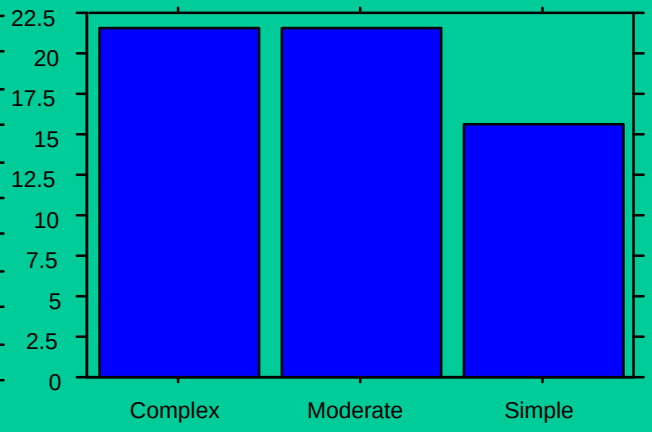


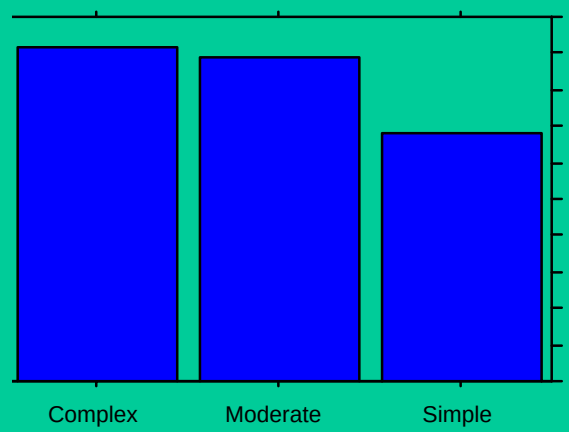
26.5 22.2 17.7

Without Ju|hoan, the result agrees with pattern found in survey of **all** languages in sample

Since 'simple' syllable-structure languages are somewhat uncommon, any areally-structured or random sample is likely to include a very small number of them. A different type of quota subsample can be formed by selecting equal numbers of the 3 syllable-structure categories







23.8 22.7 15.8 21.6 21.6 15.7 23.0 22.3 17.0

3 random selections of 10 languages in each category show consistently lower C inventory size in ‘simple’ cases

Validation

Two strategies for seeking validation of the positive correlation between syllable-structure complexity and consonant inventory size tend more to confirm than to reject the pattern found in the entire data set

A structured subsample seems the best approach to validating relationships between properties with very unequal frequencies

The most robust aspect of the syllable/consonant inventory relationship is that languages with **simple syllable structure** have a strong tendency to have **smaller than average consonant inventories**

Final Discussion

Although individual languages may historically ‘trade’ elaboration in one subsystem for simplification elsewhere, such ‘compensation’ is not an overall design feature of language; nor is there a universal pattern of co-occurring complexity

Languages vary quite considerably in their phonological complexity, as measured by the indices used here.. The ‘typical’ language is not constrained to limit complexity at this level by processing or memory constraints

Thank you for listening

References (to be completed!)

Dryer, Matthew. 1999.

Kusters, Wouter & Pieter Muysken. 2001. The complexities of arguing about complexity. *Linguistic Typology* 5: 182-186.

Maddieson, Ian. 1992. The structure of segment sequences. In *Proceedings of the 1992 International Conference on Spoken Language Processing* (Banff, Alberta), Addendum. 1-4.

Nichols, Johanna. 2003, ms. Samples for comparative grammar: A practical guide. University of California, Berkeley.

Perkins, Revere. 1989. Statistical techniques for determining language sample size. *Studies in Language* 13: 293-315.

Shosted, Ryan K. ms, 2004. Correlating complexity: a typological approach. University of California, Berkeley. (submitted to *Linguistic Typology*)

Stephens & Justeson, John. 19xx