Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts

Ian Maddieson

University of California, Berkeley

Abstract. It is often suggested that languages are likely to 'compensate' complexity in one subsystem by simplicity elsewhere. In this paper evidence against this idea is presented by examining several subsystems of the basic phonology in a set of over 600 languages selected to represent genetic and areal diversity. The relationships between elaboration of the syllable canon, the size of segment inventories and the complexity of tone systems are studied. The languages are grouped into three syllable complexity classes. The 'simple' class permits only (C)V patterns. 'Moderate' languages allow CC onsets with common structures (those with an approximant or 'liquid' as C2) and/or permit a single coda consonant. 'Complex' languages allow more elaborate clusters. Languages are also divided into three tonal groupings, those with no tone contrasts, those with simple tone systems (two levels) and those with more elaborate tone systems. Consonant, total vowel and vowel quality inventories are numeric values. There is a significant positive correlation between the complexity of syllable type and size of consonant inventory, but no correlation between syllable complexity and size of total vowel or vowel quality inventories. Consonant and vowel inventory sizes show no correlation with each other. Complex syllable structure shows some association with absence of tone, but none of the other comparisons show that complexity is systematically compensated for by simplicity elsewhere. [work supported by the National Science Foundation through grant BCS-034578.]

Keywords. Complexity, syllable structure, phonological typology

1. Introduction.

The value of casting your net wide in order to catch patterns that are widespread across languages is often demonstrated in the work of John Ohala. For example, in his paper "Southern Bantu vs the world: the case of palatalization of labials" (Ohala1978), instances of palatalized bilabials becoming coronals are assembled from Slavic, Daic, Tibetan, Romance, Greek, Germanic, Nupoid and Bantu language groups to show that this change is not an isolated oddity. Other examples could have been added from Austronesian (Maddieson 1989) and elsewhere. With the data assembled, it is then possible to interrogate the results to look for generalizations concerning the factors that favor the process and hence to gain insights into its cause. Such data-gathering and interpretation can be regarded as an 'experiment' in the natural world.

One strand of my work has also involved casting a wide net, in particular through large-scale sampling of salient phonological data from the languages of the world. In this paper a large database of phonological information is examined in order to investigate an apparently widely-held belief among linguists that complexity in one sub-part of the grammar of a language is likely to be compensated for by simplification in another. Such a view seems to be based on the humanistic principle that all languages are to be held in equal esteem and are equally capable of serving the communicative demands placed on them. In rejecting the notion of 'primitive' languages linguists seem to infer that a principle of ***equal*** complexity must apply. For example, in a widely-used basic linguistics textbook Akmajian et al (1979: 4) assert that "all known languages are at a similar level of complexity and detail — there is no such thing as a primitive human language." From this assertion it might seem a logical further step to hypothesize that languages will undergo adjustments to equalize their overall complexity across different subsystems. Thus, as far as phonology is concerned, it seems to be widely believed in the linguistic community (although such beliefs are rarely expressed in print) that, for example, a language with a large number of distinct consonants is likely to have a small number of vowels, or that a language with an elaborated tone inventory will have a tendency to compensate by having simple syllable structure.

Apart from a rather doctrinaire assertion of equal complexity, there are at least two other bases for arguing that complexity will be balanced across subsystems. Some linguists may argue that complexity in individual languages is maintained at a roughly constant overall level because that is the way historical processes work. For example, tone-splitting or tonogenesis is commonly seen as a process in which elaboration of tonal contrast is exchanged for simplification of the consonant inventory (e.g..Matisoff 1973). If it is a typical, or even relatively frequent characteristic of diachronic phonological processes that they involve exchanging complexity in different subsystems, then apparent compensatory relationships should be detectable across any reasonably inclusive sample of languages.

Others may propose that complexity is held in check because of processing considerations. Overly complex packaging of information may over-stretch our brain's ability to deal with the input in one way or another. For example, Pellegrino et al (ms, 2005) have recently suggested that the more complex syllable structure and larger inventory of English allows it to convey more information (in the sense of Shannon 1948) per syllable than does Japanese which has a simpler syllable canon and fewer distinct segments. However, English compensates for the richer information-per-syllable flow by habitual use of a slower speech rate in terms of syllables per second. Hence the two languages are seen as seeking to optimize the rate at which information is encoded and has to be decoded on-line. Again, if compensatory relationships of this general type are a required characteristic of language design, their impact should be detectable in a survey of phonological properties of languages.

Five relatively simple variables that can be taken to represent certain aspects of the complexity of phonological systems will be compared pairwise to see whether they tend to correlate positively (greater complexity going together on both measures) or negatively (greater complexity on one measure going with lower complexity on the other), or are simply uncorrelated. The variables describe the degree of elaboration of the syllable canon, the size of inventories of consonants, basic vowel qualities, and all vowels and the complexity of tone systems. The sample of languages examined will first be described, together with how these phonological characteristics are represented in the data.

2. Language sample and data

The language sample is essentially an expanded version of the UPSID sample described in Maddieson (1984). This was earlier enlarged from 317 to 451 languages for a version distributed as an interactive database (Maddieson & Precoda 1990). At both these stages, selection of languages for inclusion was governed by a quota principle seeking maximum genetic diversity among extant languages (and those spoken until recently). More recently, the UPSID sample was merged with the 200 language list chosen as the core sample for the *World Atlas of Language Structures* (WALS) (Haspelmath et al 2005). Although representation of genetic diversity was a major goal of this sample, a number of other factors also played a role, including the political importance of a language, the availability of a full-length grammar, the presence of the language in other typological samples and a desire for wide geographical coverage,. Due to overlapping membership of the samples this resulted in a merged list of about 520 languages. The languages previously included in a study of syllable structure (Maddieson 1992) or figuring in a long-term project to investigate the phonetic characteristics of endangered languages have also been added. As part of a project to more fully examine geographical distribution of phonological characteristics of languages, this database is continuing to be enlarged, with the intention of reaching a total of 1000 or so languages. It should be noted that the addition of more languages, and in particular the relaxation of the requirement that no pair of closely related languages be included, changes the representative nature of the sample from that aimed for in the original UPSID sample.

At the time of writing the database includes 625 languages, with the data for different languages at varying levels of completeness. As shown in Table 1, there are approaching or over one hundred languages in each of the six major areal/genetic groupings used for subsetting the data for purposes of statistical validation. The composition of these groupings, which are broadly similar to those used by Dryer (1992, 2003), is described in detail in Maddieson (2005). In forming these groupings genetic factors override purely areal ones in the following way. In order to keep related languages together all languages in a family are included in the area of its major concentration. For example, Semitic languages spoken in Asia Minor are grouped with other Afro-Asiatic languages under the African area, and all Austronesian languages are grouped together in the East and South-East Asian area, including even Malagasy and Maori. The table is included here primarily to confirm the global distribution of the languages in the sample, but occasional reference to differences between the language groups will be made.

Table 1. Distribution of languages by areal/genetic grouping

Area	**# of languages in sample**
Europe, West & South Asia	94
East & South-East Asia	119
Africa	139
North America	91
South & Central America	87
Australia & New Guinea	95

The first property to be examined is the complexity of the maximal syllable structure the language permits. In order to reduce a considerable variety of patterns to a manageable number of categories, the languages are divided into Simple, Moderate and Complex categories based on what is permitted in syllable onset and coda positions. Those with Simple structure allow no more than one onset consonant and do not permit codas. In other words they permit only (C)V syllables. Among languages in this category are Yoruba, Jul'hoan, Maori and Guarani. Languages with a Moderate level of syllabic complexity are those in which the coda may contain not more than one consonant, and the onsets are limited to a single consonant or a restricted set of two-consonant clusters having the most common structural patterns, typically an obstruent followed by a glide or liquid (the 30 language sample used in Maddieson (1992) had shown that simple codas and minimally elaborated onsets tend to co-occur in languages). Among languages in this class are Nahuatl, Tigre, Yanyuwa and Mandarin (the "medials" of the traditional analysis of Chinese syllables are treated as glides in an onset cluster). Languages which permit a wide range of onset clusters, or which permit clusters in the coda position are classified as belonging to the Complex syllable category. Languages in this class include Georgian, Quileute, Soqotri and French. No account is taken of the relative frequency of the different syllable patterns in the lexicon, and patterns restricted to relatively recent loanwords are ignored. Values for the syllable complexity category have been entered so far for 564 of the 625 languages, with a large preponderance belonging to the Moderate category (320 or 56.7%), and Complex languages outweighing Simple ones among the remainder by 180 to 64 (31.9% to 11.3%).

The complexity of the tone system is also reduced to a three-way categorical variable. Languages with no tonal contrasts form the first group, those with a tone system that can be reduced to a basic two-way contrast (usually of level high and low tones) form the second group, and those with a more complex tone system containing three or more levels or contours form the third. A number of languages which do not have canonical tone systems but where pitch patterns make basic lexical distinctions, e.g. Japanese and Norwegian, have been grouped with the Simple tone languages, as have a few languages with recognized but under-analyzed tone systems. Tone category has been entered for 572 of the 625 languages, with a majority (338 or 59.1%) having no tones, 140 or 24.5% having Simple tone systems (including 14 languages where the tone system has the marginal status mentioned above), and 94 or 16.4% have Complex tone systems. Complex tone systems occur far more frequently in the East and South-East Asian and African groups than elsewhere, but Simple tone systems and languages with no tone system are distributed fairly uniformly across the areal/genetic groupings outside Africa. In the African group a majority of languages have a Simple tone system and languages without tone are few in number.

The remaining three properties compared are numerical rather than categorical. The consonant inventory size is the total number of distinctive consonants recognized for the language in a phonemic-style analysis. The vowel quality inventory size is the number of distinct vowel types contrasting on the major parameters of height, rounding and backness, i.e. independent of length, nasalization, voice quality, or other 'series-generating' components. Distinctions that are known or hypothesized to be based on tongue root position are equated with height distinctions, in part because for many languages the role played by tongue root is not known. The total vowel inventory size is the number of distinct vowel nuclei including distinctions of nasality, phonation type, and length as well as any nuclear diphthongs (i.e. those which do not consist of a vowel and a consonantal glide or a sequence of two separate vowels). An attempt has been made to reach interpretations of the consonant and vowel inventories based on uniform principles, so in many cases they depart from the analysis found in published descriptions of the individual languages. Undoubtedly there are errors of judgement or fact affecting the interpretations made, but in a large sample errors in different directions can be expected to cancel each other out, and any robust general patterns can be expected still to emerge.

Consonant inventory and vowel quality inventory values are currently entered for 617 of the 625 languages. Consonant inventory ranges from 6-128 with a mean of 22.6, and vowel quality inventory from 2-16 with a mean of 6.0. Total vowel inventory is entered for only 535 languages (questions about the status of vowel length and potential diphthongs often being as yet unresolved), and ranges from 2-46 with a mean of 10.2. Because a small number of very salient outliers contribute a serious distortion to numerical analyses, languages with more than 30 total vowels or 70 or more consonants are discarded when these variables are considered. Each of these thresholds is exceeded by three languages, one of which, Jul'hoan, exceeds both.

3. Relationships between variables

In this section, each of the five variables described in the previous section will be compared pairwise with the others, except for the vowel quality and total vowel inventories (which are evidently correlated). When a categorical and a numeical variable are compared, the means of the numerical variable in each category form the basis of the comparison. When two numerical categories are compared a simple regression is used. The comparison of syllable and tone categories is made using an index. The number of languages in each comparison varies according to the number with values specified for both variables. In each case, the purpose is to see if greater complexity on the two variables tends to co-occur, or if a compensatory relationship exists, or there is no overall trend of either kind.

In the first three comparisons syllable structure is compared with the segment inventory variables. As figure 1 shows, syllable structure complexity and the size of the consonant inventory are positively correlated. The mean number of consonants in the inventory is greater for each increase in complexity of the maximal syllable. Analysis of variance show a highly significant effect of syllable category on consonant inventory size (F (2, 556) = 23.26, p < .0001) and all pairwise comparisons are highly significant in a post-hoc comparison (using Fisher's PLSD adjusted for unequal cell sizes). In Figure 1 and subsequent figures the error bars show the 95% confidence interval.

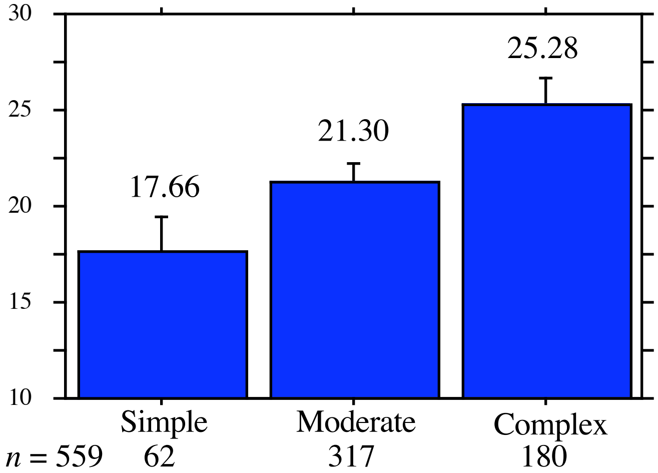

Figure 1. Mean number of consonants by syllable category.

Neither of the two measures of vowel inventory shows a systematic relationship with syllable complexity. In each case the mean value within each syllable category is essentially the same as the grand mean. These comparisons are shown in Figures 2 and 3. Neither displays a significant effect for syllable category (F (2, 566) = .216, p = .8054 for vowel quality and F(2, 488) = 0.278, p = .757 for total vowel inventory).

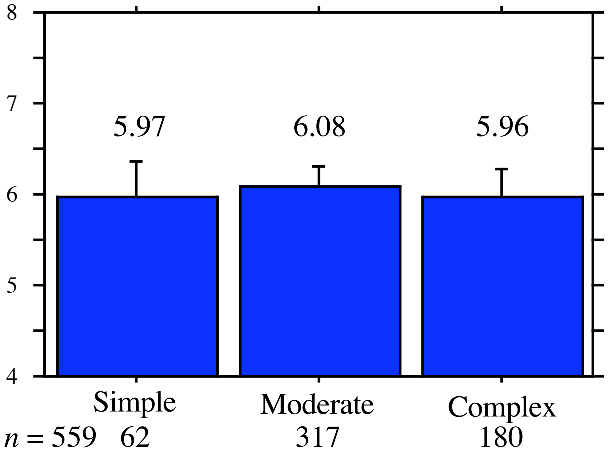

Figure 2. Mean number of vowel qualities by syllable category.

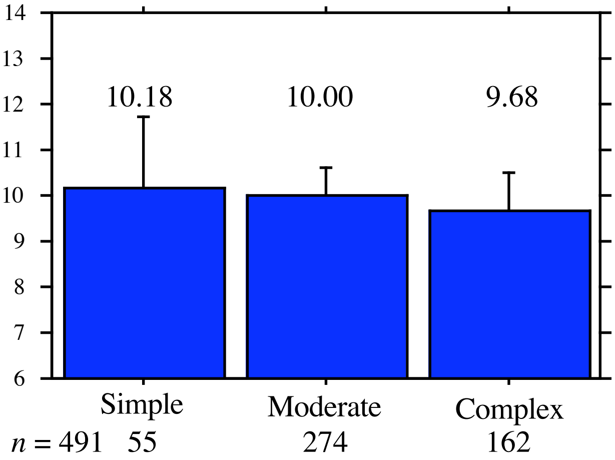

Figure 3. Mean number of total vowels by syllable category.

The size of the consonant inventory correlates positively (though rather weakly) with increasing complexity of tone system, as shown in Figure 4. Analysis of variance shows a significant effect of tone category at better than p < .05 (F (2, 566) = 3.336, p < .0363). Posthoc comparisons show that only the comparison between 'None' and 'Complex' reaches significance.

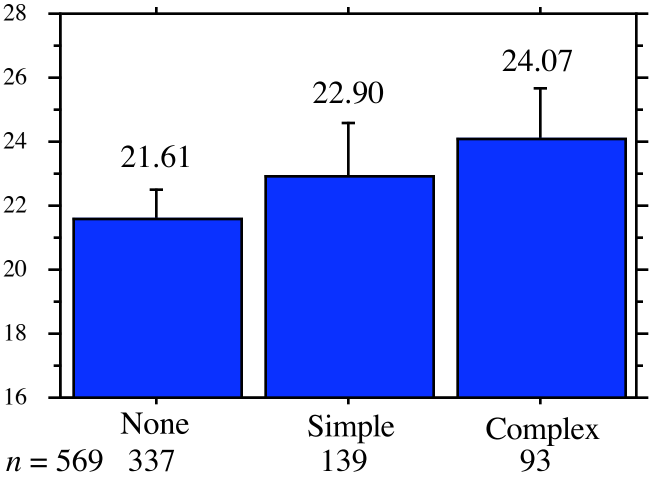

Figure 4. Mean number of consonants by tone category.

The size of the vowel quality inventory and the complexity of tone system also positively correlate with each other; as tone complexity increases, so does the mean number of distinct vowel qualities. This result is shown in Figure 5. In the analysis of variance there is a highly significant effect of tone cateogory (F (2, 566) = 20.591, p < .0001) and all posthoc pairwise comparisons are significant at better than p < .05, with the difference between 'None' and either category of tonal languages better than p < .001.

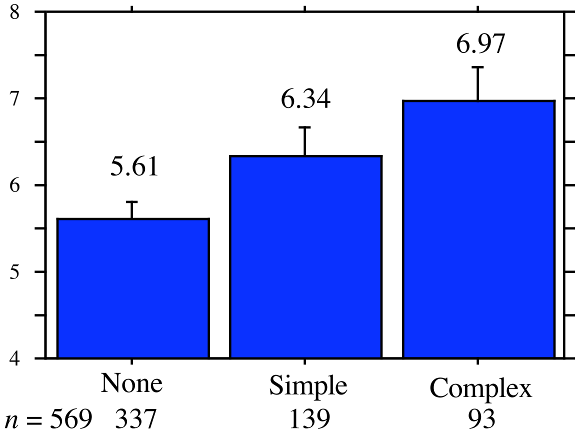

Figure 5. Mean vowel quality inventory by tone category.

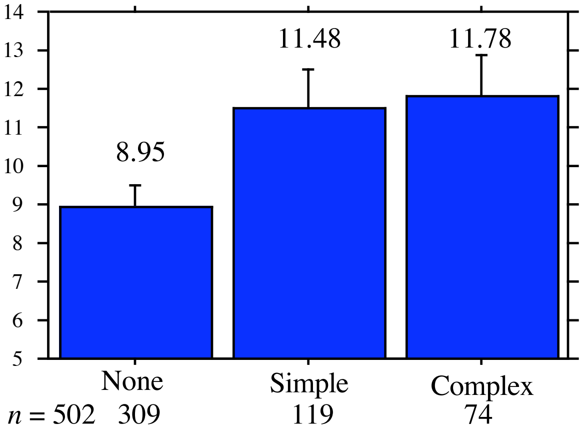

Figure 6. Mean total vowel inventory by tone category.

There is also a correlation between an increase in total vowel inventory size and the presence of a tone system, as shown in Figure 6. There is a highly significant overall effect of tone category in the analysis of variance (F (2, 499) = 16.554, p < .0001) which the posthoc comparisons indicate is due to a highly significant difference between 'None' and either tonal category, with no significant difference being found between the two categories of tonal languages.

There is no systematic relationship between the number of vowel qualities and the number of consonants in the inventories of the languages (N = 612) , nor between the total number of vowels and the number of consonants (N = 530). Regression plots are shown in Figures 7 and 8 for these two comparisons. For the regesssion in Figure 7 the R^2 value is .0002, and for that in Figure 8 it is .002.

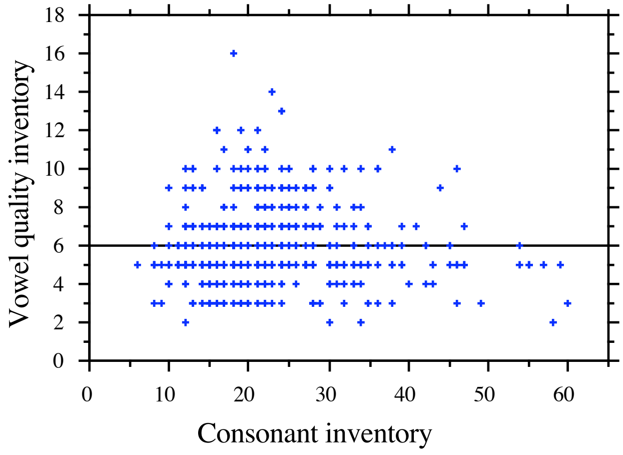

Figure 7. Regression plot of consonant inventory and vowel quality inventory.

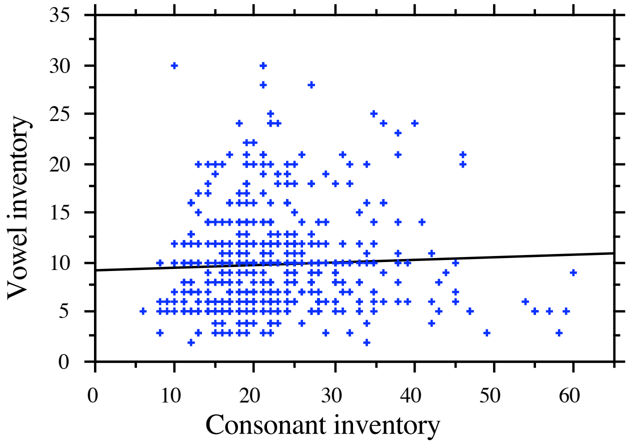

Figure 8. Regression plot of consonant inventory and total vowel inventory.

The final comparison is between the two categorical variables reflecting complexity of syllable structure and tone system. Tone system complexity does not associate with the complexity of syllable structure; rather the occurrence of complex syllable structure and lower tonal complexity are associated. In the total sample of 543 languages examined for this relationship, 88 or 16.2 % have complex tone systems, but among the 172 languages with complex syllable structure only 11 or 6.4% have a complex tone system. Another way to illustrate this pattern is with a tonal complexity index. Languages with no tones are coded with 1, languages with a simple tonal system as 2, and those with a complex tonal system as 3. The mean value of this index is then computed for each syllable complexity class. Results are shown in Figure 9. There is a significant overall effect (F (2, 540) = 19.15, p < .0001), with the index being significantly lower for the complex syllable structure class than for the other two categories, which are not significantly different in posthoc comparisons.

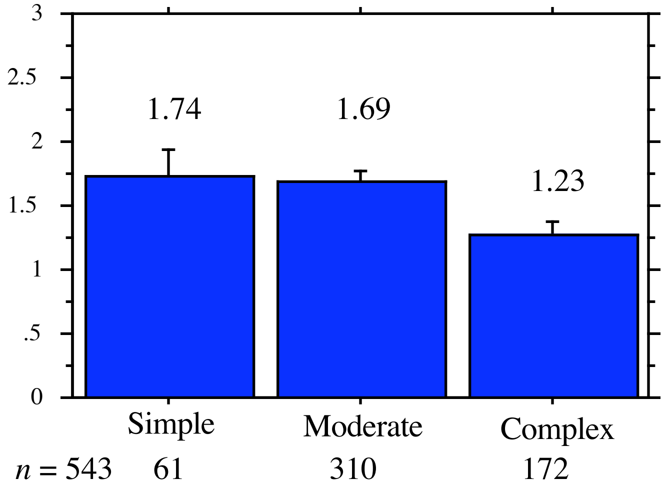

Figure 9. Tone complexity index by syllable category.

Summary and discussion of results

The nine comparisons made in the preceding section show almost no evidence of any tendency for languages to 'compensate' for increased complexity in one of the phonological subsystems surveyed by greater simplicity in another. In fact four of the comparisons show positive correlations of increases in complexity between different subsystems. Increasing syllabic complexity is positively associated with increasing size of consonant inventory, and increasing complexity of tone system is positively associated with increasing size of both consonant and vowel inventories. The mean number of both vowel qualities and total vowels increases with tone complexity. Although these two quantities are generally associated with each other, the two relationships should be considered separately. It would be quite possible for the 'series-generating' components which are mainly responsible for the differences between the two numbers to be distributed across the tone categories in a way that created a compensatory relationship between total vowel inventory and tonal complexity.

A further four of the comparisons show no systematic relationship at all between the variables examined. These are the two relationships between syllable complexity and the vowel measures, and the two relationships between consonant inventory size and the vowel measures.

Despite anecdotally-based belief to the contrary, increasing elaboration of consonant inventory is unrelated to size of vowel quantity inventory (as earlier noted by Stephens &. Justeson 1984).

Only one of the relationships shows a tendency to show compensation — that between syllable complexity and tone. Languages with complex syllable structure are most frequent in the Europe, West & South Asia, and North American language groups, and these are two of the four areas in which relatively few of the languages are tonal.

Although individual languages may historically 'trade' elaboration in one subsystem for simplification elsewhere, such 'compensation' does not seem to be a design feature of language. Nor do we find here evidence that languages are shaped in a compensatory fashion to avoid testing the limits of processing ability. It simply seems to be the case that languages vary quite considerably in their phonological complexity, as measured by the indices used here. Similar variability and absence of compensatory patterning was also found by Shosted (2004) in comparing phonological and morphosyntactic elaboration.

It should be noted that the results in this paper are presented as descriptive statistical summaries of the data that is entered into the database. The design and use of language surveys in typological studies and universals research and the nature of the inferences that can fairly be drawn from their analysis have been topics of considerable interest (see, for example, Dryer 1989, Perkins 1989, 2001, Cysouw to appear). Much concern centers on the question of whether the languages included in a sample can be considered independent. The concern is particularly acute for those who hope to be able to extrapolate from a sample of documented languages to the universe of possible languages. I do not think this step can ever be justified. Moreover, when the hypothesis for which a sample is being used concerns compensatory adjustments, it is not clear that closeness in genetic or areal terms to an included language should disqualify the inclusion of another. After all, in the kind of historical scenario that forms one foundation for the expectation of compensatory adjustments, comparison of dialects of the same language can provide the model. For example Northern Khmu? has tone but has neutralized obstruent voicing, Southern Khmu? has no tone and an obstruent voicing distinction (Svantesson 1983). The issue is whether changes that create such patterns occur often enough to leave an overall imprint on language structures. The best way to search for the answer may well be to examine the largest and most inclusive sample possible.

References

Akmajian, Andrew, Richard A. Demers, & Robert M. Harnish. 1979. *Linguistics: An Introduction to Language and Communication.* Cambridge, MA: MIT Press

Cysouw, Michael. to appear. Quantitative methods in typology. In: Gabriel Altmann, Reinhard Köhler, R. Piotrowski (Eds.).*Quantitative Linguistik - Quantitative Linguistics: An International Handbook.* (HSK, Band 27). Berlin: Mouton de Gruyter.

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

Dryer, Matthew. 1992. The Greenbergian word order correlations. *Language* 68: 81-138.

Dryer, Matthew S. 2003. Significant and non-significant implicational universals. *Linguistic Typology* 7: 108-12

Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie, eds. 2005. *The World Atlas of Language Structures.* Oxford and New York: Oxford University Press.

Maddieson, Ian. 1992. The structure of segment sequences. In J. J.Ohala et al, eds *Proceedings, 2nd International Conference on Spoken Language Processing*, Addendum. 1-4.

Maddieson, Ian. ms, 2005 Correlating phonological complexity: data and validation. Submitted to *Linguistic Typology*.

Maddieson, Ian and Kristin Precoda. 1990. Updating UPSID. *UCLA Working Papers in Phonetics* 74: 104-114.

Maddieson, Ian. 1984. *Patterns of Sounds*. Cambridge: Cambridge University Press.

Maddieson, Ian. 1989. "Linguo-labials". In R. Harlow & R. Hooper, eds, *VICAL, Papers from the Fifth International Conference on Austronesian Linguistics, Vol I, Oceanic Languages*, 349-375. Auckland: Linguistic Society of New Zealand.

Matisoff, James. 1973. Tonogenesis in South-East Asia. In L. M. Hyman, ed, *Consonant Types and Tone*: 71-95. Los Angeles: University of Southern California

Ohala, John. 1978. Southern Bantu vs the world: the case of palatalization of labials. Berkeley Linguistic Society 4:

Pellegrino, François, Christophe Coupé & Egidio Marsico . ms, 2005. Cross-Linguistic Comparison of Phonological Information Rate. Laboratoire Dynmaique du Langage, Université Lyon-2.

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

Perkins, Revere 2001. Sampling procedures and statistical methods. In M. Haspelmath et al, eds, *Language Typology and Language Universals: An International Handbook* (HSK, Band 20): 419-434. Berlin: Walter de Gruyter.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379-423, 623-656.

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

Stephens, Laurence D. & John S. Justeson 1984. On the relationship between numbers of vowels and consonants in phonological systems. *Linguistics* 22: 531-545.

Svantesson, Jan-Olof. 1983. Kammu phonology and morphology. (Travaux de L'Institut de Linguistique de Lund, 18.) Lund: Gleerup.