Preface

The discovery of generalizations concerning the content and structure of phonological inventories has been a significant objective of recent work in linguistics. Such generalizations have been taken into account, explicitly or implicitly, in the formulation of phonological theories, in evaluating competing historical reconstructions, in constructing models of language change and language acquisition, and they have stimulated important linguistically-oriented phonetic research. This book reports on the work done at UCLA using a computer-accessible database containing the phonological segment inventories of a representative sample of the world's languages which is designed to provide a reliable basis for such generalizations. The project has come to be referred to by the acronym UPSID - the UCLA Phonological Segment Inventory Database.

There seem to be three types of sources for observations on phonological inventories. The type with the longest tradition is an essentially impressionistic account based on a linguist's experience of a number of languages. Statements by Trubetskoy (1939), Jakobson and Halle (1956), and Ladefoged (1971) as well as incidental remarks in the papers of numerous authors are examples of this category. Although they may be based on familiarity with a very large number of languages, there is some doubt about the scope and validity of the conclusions reached, since the list of languages represented in this experience is not given and there is no quantification attached to the statements made.

The second type consists of explicit samples of languages compiled for the purpose of a single study, such as Ferguson (1963), Greenberg (1970) and Hyman (1977) on nasals, glottalic consonants and stress respectively. In these cases the quality of the sample (cf. Bell 1978) and the significance of the conclusions reached (cf. Hurford 1977) can be independently assessed by the reader.

The third kind of data source is a standardized multi-purpose survey, epitomized by the Stanford Phonology Archive (SPA), compiled at Stanford University as part of the broad Language Universals Project under the direction of J. H. Greenberg and C. A. Ferguson. A large proportion of recent work on phonological universals is either directly based on the SPA or owes an indirect debt to it. The UCLA Phonological Segment Inventory Database (UPSID) is a source of this third kind.

There are several reasons for the superiority of this third kind of data source which arise from the nature of the field of enquiry involved. The data source serves, first, to generate observations, e.g. observations concerning the frequency of segments of different types and of the phonetic attributes of segments, as well as their co-occurrence in phonological inventories, and secondly, to subject hypotheses concerning such matters as segment frequency to the test of comparison with empirical observations. The hypotheses may range from simple ones claiming that there are significant differences in the frequency of segments of different types to more elaborate ones positing contingent relationships between the occurrence of (sets of) different segments, or limitations on the distribution of phonetic attributes within inventories. The third, and perhaps most significant, purpose behind compilation of such data sources is as a stimulus to the generation of hypotheses which relate to other fields of the study of language but for which such matters as segment frequencies, inventory size, and so on, may be the point of departure. Such hypotheses can be directed at issues of production, perception, acquisition, linguistic change or language contact, but establish connections between other data and observations concerning segments and inventories.

Most of these observations and hypotheses about phonological universals necessarily concern relative rather than absolute matters. Experience has shown that few interesting things are to be said about phonological inventories that are truly universal, i.e. exceptionless. Apart from observations such as "all languages have a contrast between consonants and vowels" most of the substantive generalizations concerning segments and inventories are or can be expected to be of the form "a situation x occurs more (or less) frequently than chance leads us to predict." That is, in layman's terms, they are statistical observations. They can therefore only be meaningful if they are drawn from, or tested with respect to, a body of data appropriately designed for statistical analysis. In other words, one which is representative, extensive and uniform in analysis as far as possible. This requires establishment of a large and appropriately selected sample of languages and a standardized procedure for interpreting their phonologies. Once such a database has been established, numerous commensurate studies on the same data can be made.

This book contains nine chapters presenting analyses of aspects of the UPSID inventories. Chapter 9 is contributed by Sandra F. Disner, the rest are written by me. Each of these chapters is designed to be largely self-contained so that readers may consult a single chapter if, for example, they are interested in some particular segment type. Chapter 10 presents a relatively full account of the design of the database, including the principles governing the selection of languages, the criteria used in interpretation of the descriptive sources consulted and the set of phonetic features used to characterize segments. A full documentation of the data itself is also contained in the appendices at the end of the book, including phonemic charts of each language and full lists of the types of segments that occur. Each language is assigned an identification number which is cited whenever the language is mentioned in the text, enabling the corresponding phoneme chart to be easily found. The principles on which the identification numbers are assigned is explained in Appendix A.

Many people have assisted in making this book possible. The principal work of establishing the computer database was done by Sandra F. Disner, Vivian Flores, J. Forrest Fordyce, Jonas N. A. Nartey, Diane G. Ridley, Vincent van Heuven and myself. Help in collecting data was provided by Stephen R. Anderson, Peter Austin, Steve Franks, Bonnie Glover, Peter Ladefoged, Mona Lindau-Webb, Robert Thurman, Alan Timberlake, Anne Wingate, Andreas Wittenstein and Eric Zee. Additional assistance has come from other linguists at UCLA and elsewhere. Mel Widawsky provided valuable services in persuading the computer to accept the indigestible bulk of our input. A library of the sources from which data was drawn was compiled with assistance from Hector Javkin and Diane G. Ridley. John Crothers provided an early copy of the final report of the Stanford Phonology Archive, enabling the UCLA project to benefit from the experience accrued at Stanford. Geoffrey Lindsey and Karen Weiss did the tedious work of typing the phoneme charts and Karen Emmorey, Karen Weiss, Alice Anderton, and Kristin Precoda assisted with the preparation of the camera-ready copy of the remainder of the book. To all of these people I owe an enormous debt, which I can only pay in the coin of gratitude.

I also owe thanks to those who have shown faith in the UPSID project as it developed by making use of it, including Louis Goldstein, Pat Keating, Peter Ladefoged, Björn Lindblom and the students in Linguistics 103 at UCLA.

A considerable portion of the work reported in this book has been funded by the National Science Foundation through grants BNS 78-07680 and BNS 80-23110 (Peter Ladefoged, principal investigator). Neither the NSF nor any of the individuals named above are responsible for the errors that undoubtedly remain. If you the reader find one, please write and tell me about it.

Ian Maddieson
University of California
Los Angeles

References

Bell, A. 1978. Language samples. In J.H. Greenberg et al. (eds.) *Universals of Human Language, Vol 1, Method and Theory*. Stanford University Press, Stanford: 123-56.

Ferguson, C. A. 1963. Some assumptions about nasals. In J.H. Greenberg (ed.) *Universals of Language*. MIT Press, Cambridge: 42-7.

Greenberg, J. H. 1970. Some generalizations concerning glottalic consonants, especially implosives. *International Journal of American Linguistics* 36: 123-45.

Hurford, J. R. 1977. The significance of linguistic generalizations. *Language* 53: 574-620.

Hyman, L. M. 1977. On the nature of linguistic stress. In L.M. Hyman (ed.) *Studies on Stress and Accent*. (Southern California Occasional Papers in Linguistics 4) University of Southern California, Los Angeles: 37-82.

Jakobson, R. and Halle, M. 1956. Phonology and phonetics. (Part 1 of) *Fundamentals of Language*. Mouton, The Hague: 3-51.

Ladefoged, P. 1971. *Preliminaries to Linguistic Phonetics*. University of Chicago Press, Chicago.

Trubetskoy, N. 1939. *Grundzüge der Phonologie* (Travaux du Cercle Linguistique de Prague 9). Prague.

1

The size and structure of phonological inventories

1.1 Introduction

A database designed to give more reliable and more readily available answers to questions concerning the distribution of phonological segments in the world's languages has been created as part of the research program of the UCLA Phonetics Laboratory. The database is known formally as the UCLA Phonological Segment Inventory Database, and for convenience is referred to by the acronym UPSID. UPSID has been used to investigate a number of hypothesized phonological universals and "universal tendencies". Principal among these have been certain ideas concerning the overall size and structure of the phonological inventories. The design of the database is briefly described in this chapter. A full description is given in chapter 10, and the various appendices at the end of the book report on the data contained in UPSID files. The remainder of the present chapter discusses the issues involving the overall structure and size of phonological inventories which have been examined with its use.

1.2 Design of the database

The languages included in UPSID have been chosen to approximate a properly constructed quota sample on a genetic basis of the world's extant languages. The quota rule is that only one language may be included from each small family grouping, for example, among the Germanic languages, one is included from West Germanic and one from North Germanic (East Germanic, being extinct and insufficiently documented for a reliable phonological analysis to be made, is not included). Each such small family grouping should be represented by the inclusion of one language. Availability and quality of phonological descriptions are factors in determining which language to include from within a group, but such factors as the number of speakers and the phonological peculiarity of the language are not considered. The database includes the inventories of 317 languages. In this and subsequent chapters, every language mentioned in the text is identified by a number that cross-refers to the list of these languages and the data charts at the end of the book. These numbers are assigned on the basis of the genetic affiliation of the language.

In the database each segment which is considered phonemic is represented by its most characteristic allophone, specified in terms of a set of 58 phonetic attributes. These are treated as variables which take the value 1 if the segment has the attribute and 0 if the segment lacks it. The list of attributes with the value 1 thus provides a phonetic description of the segment concerned.

For 192 of the 317 languages included, UPSID has profited from the work of the Stanford Phonology Archive (SPA). Our decisions on phonemic status and phonetic description do not always coincide with the decisions reached by the compilers of the SPA, and we have sometimes examined additional or alternative sources, but a great deal of effort was saved by the availability of this source of standardized analyses. It should be noted that UPSID, unlike the SPA, makes no attempt to include information on allophonic variation, syllable structure, or phonological rules.

In determining the segment inventories, there are two especially problematical areas. The first involves choosing between a unit or sequence interpretation of, for example, affricates, prenasalized stops, long (geminate) consonants and vowels, diphthongs, labialized consonants, etc. The available evidence which bears on the choice in each language individually has been examined but with some prejudice in favor of treating complex phonetic events as sequences (i.e. as combinations of more elementary units). The second problem area involves the choice between a segmental and a suprasegmental analysis of certain properties. Stress and tone have always been treated as suprasegmental; that is, tonal and stress contrasts do not by themselves add to the number of distinct segments in the inventory of a language, but if differences in segments are found which accompany stress or tone differences, these may be regarded as segmental contrasts if the association does not seem a particularly natural one. For example, if there is an unstressed vowel which is a little shorter or more centralized than what can be seen as its stressed counterpart, these vowels will be treated as variants of the same segment. However, larger qualitative differences between the set of stressed and unstressed vowels will lead us to enter such sets of vowels as separate segments. In all cases, sets of vowels which are divided into vowel harmony series are all entered separately; the factor which distinguishes the vowel harmony series is not extracted as a suprasegmental.

1.3 Variations in inventory size

The number of segments in a language may vary widely. The smallest inventories included in the survey have only 11 segments (Rotokas, 625; Mura, 802) and the largest has 141 (!Xũ, 918). However, it is clear that the typical size of an inventory lies between 20 and 37 segments - 70% of the languages in the survey fall within these limits. The mean number of segments per language is a little over 31; the median falls between 28 and 29. These values are very close to the number 27 ± 7 which Hockett (1955) estimated as the most likely number of segments in a language.

The variability in segment totals can be reflected in a number of statistical measures. These show that the curve formed by plotting the number of languages against the segment totals is not normally distributed. It is both positively skewed and platykurtic, that is, there is a longer tail to the distribution at the high end of the scale, and the shape of the curve is one with a low peak and heavy tails. This implies that the mean number of segments is not a good way to sum up the distribution. For this reason more attention should be paid to the range 20-37 than the mean of 31.

Whether the tendency to have from 20 to 37 segments means that this is an optimum range is an open question. It seems likely that there is an upper limit on the number of segments which can be efficiently distinguished in speech, and a lower limit set by the minimum number of segments required to build an adequate vocabulary of distinct morphemes. But these limits would appear to lie above and below the numbers 37 and 20 respectively.

Consider the following: the Khoisan language !Xũ (918) with 141 segments is related to languages which also have unusually large inventories. Comparative study of these languages (Baucom 1974; Traill 1978) indicates that large inventories have been a stable feature which has persisted for a long time in the Khoisan family. If the number of efficiently distinguished segments was substantially smaller, there would be constant pressure to reduce the number of segments. There does not seem to be any evidence of such pressure.[1]

Similarly, the facts do not seem to show that languages with small inventories (under 20 segments) suffer from problems due to lack of contrastive possibilities at the morphemic level. The symptoms of such difficulties would include unacceptably high incidence of homophony or unmanageably long morphemes. Dictionaries and vocabularies of several languages with small inventories, such as Rotokas (625, Firchow, Firchow and Akoitai 1973), Hawaiian (424, Pukui and Elbert 1965) and Asmat (601, Voorhoeve 1965: 293-361), do not provide evidence that there are symptoms of stress of these kinds in languages with small phoneme inventories. Hawaiian, for example, with 13 segments has been calculated to have an average of just 3.5 phonemes per morpheme (Pukui and Elbert 1965: xix), clearly not unacceptably long. And again, comparative evidence indicates that small inventory size may be a phenomenon which persists over time, as, for example, in the Polynesian language family, which includes Hawaiian (Grace 1959).

The restrictions on inventory size may therefore not be theoretical ones relating to message density and channel capacity in language processing. Although such considerations have been the most widely discussed, they are far from the only ones likely to influence the typical language inventory. Linguistic messages do have to be sufficiently varied to be able to deal with myriad situations and they do need to be successfully conveyed via a noisy channel, but the design of language is also subject to many pressures of a "non-functional" kind. Most languages exist in a multi-lingual social context. Limits may be placed on the size of a typical inventory through language contact, especially situations where a language is gaining speakers who are learning the language after early childhood. The mechanism may be one which approximates the following: speakers acquiring a new language make substitutions for any segment that is not matched by a closely similar segment in their own language, or is not capable of being generated by a simple process of adding familiar features (e.g. acquiring /g/ is easy if you already have /p, b, t, d / and /k/ in the first language). The resulting inventory in the acquired language contains only the segments common to both input languages, plus a few segments "generated" by the process outlined above. The smaller the inventory of the first language, the greater the probability that some segments will be generated in the fashion outlined. The greater the inventory, the smaller the probability that similar segments will coincide in the two languages and thus the greater the probability of inventory simplification.

This proposal predicts not only that upper and lower limits on inventory size will tend to be rather flexible, as is the case, but also that areal-genetic deviations from the central tendency should be expected. Thus, greater than average size inventories in Khoisan or Caucasian languages, and smaller than average in Polynesian are understandable results: local deviations are perpetuated because primary contact is with other languages tending in the same direction. This proposal also avoids a difficulty; if human processing limitations are postulated as the cause of limitations on the size of inventories, then they ought invariably to exert pressure to conform on the deviant cases. The evidence for this is lacking.

1.4 Relationship between size and structure

The data in UPSID have been used to address the question of the relationship between the size of an inventory and its membership. The total number of consonants in an inventory varies between 6 and 95 with a mean of 22.8. The total number of vowels varies between 3 and 46 with a mean of 8.7. The balance between consonants and vowels within an inventory was calculated by dividing the number of vowels by the number of consonants. The resulting ratio varies between 0.065 and 1.308 with a mean of 0.402. The median value of this vowel ratio is about 0.36; in other words, the typical language has less than half as many vowels as it has consonants. There are two important trends to observe; larger inventories tend to be more consonant-dominated, but there is also a tendency for the absolute number of vowels to be larger in the languages with larger inventories. The first is shown by the fact that the vowel *ratio* is inversely correlated with the number of consonants in an inventory ($r = -.40$, $p = .0001$) and the second by the fact that the *total* of vowels is positively correlated with the consonant total ($r = .38$, $p = .0001$). However, a large consonant inventory with a small vowel inventory is certainly possible, as, for example, in Haida (700: 46C, 3V), Jaqaru (820: 38C, 3V) or Burushaski (915: 38C, 5V). Small consonant inventories with a large number of vowels seem the least likely to occur (cf. the findings of Hockett 1955), although there is something of an areal/genetic tendency in this direction in New Guinea languages such as Pawaian (612: 10C, 12V), Daribi (616: 13C, 10V) and Fasu (617: 11C, 10V). In these cases a small number of consonants is combined with a contrast of vowel nasality. Despite some aberrant cases, however, there is a general though weak association between overall inventory size and consonant/vowel balance: larger inventories tend to have a greater proportion of consonants.

Such an association suggests that inventory size and structure may be related in other ways as well. A simple form of such a hypothesis would propose that segment inventories are structured so that the smallest inventories contain the most frequent segments, and as the size of the inventory increases, segments are added in descending order of their overall frequency of occurrence. If this were so, all segments could be arranged in a single hierarchy. Such an extreme formulation is not correct, since no single segment is found in all languages. But if we add a corollary, that larger inventories tend to exclude some of the most common segments, then there is an interesting set of predictions to investigate. We may formulate these more cautiously in the following way: a smaller inventory has a greater probability of including a given common segment than a larger one, and a larger inventory has a greater probability of including an unusual segment type than a smaller one.

The extent to which languages conform to the predictions can be tested in two straightforward ways. One is to examine inventories of some given size and see what segments they contain; the other is to examine given segment types and see how they are distributed across inventories by size. Using the second approach, the distribution of 13 of the most frequent consonants was investigated in a set of UPSID languages with relatively small inventories and in a set of languages with relatively large inventories. For the small inventory set, languages with 20-24 segments were chosen. Below 20 segments a language usually has fewer than 13 consonants, so that exclusions would occur simply because of the small numbers involved. For the large inventory set, all UPSID languages with over 40 segments were selected. These choices resulted in subsamples containing 57 and 54 languages respectively.

The set of consonants investigated and their distribution is shown in Table 1.1 below, together with three percentages. The first is the percentage of the 57 small inventory languages with the given segment, the second is the percentage of all UPSID languages which have the segment and the third is the percentage of the large inventory languages which have the segment. Note that consonants in the dental/alveolar region have not been considered here because of the frequent uncertainty as to whether they are dental or alveolar.

The consonants investigated fall into three groups. Using the overall frequency of the segment as the expected value, the first and third groups of these consonants show significant deviations ($p < .005$), while the