CONSTRUCTING A GLOBAL CROSS-LINGUISTIC DATABASE OF BASIC PHONOLOGICAL PROPERTIES: PRINCIPLES AND CHALLENGES

Ian Maddieson

University of New Mexico and University of California, Berkeley
ianm@berkeley.edu

ABSTRACT

There is an obvious interest in capturing general trends in the structure of phonological systems of the world's extant languages. These may hint at overall design properties of human language, which in turn may have origins in basic human cognitive properties or characteristics inherited from the earliest human language(s). One tool that can be used to study such trends is a broadly-based cross-linguistic database on phonological systems. Four of the principal challenges to providing this will be discussed in this paper, which describes the thinking behind the compilation of the LAPSyD database and draws some comparisons with other somewhat similar projects, such as PHOIBLE, SAPhon and Segérer's African consonant inventory database.

Keywords: phonological inventories, databases

1. INTRODUCTION

As with other typological traits there is an obvious interest in capturing general trends in the structure of phonological systems of the world's extant languages. Overall design properties of language, in turn having origins in basic human cognitive abilities or reflecting inherited characteristics of the first human language(s) can possibly be inferred from knowledge of such patterns. One tool to study such trends is a broadly-based cross-linguistic database on phonological systems. There are several challenges to designing such a project. Four of the principal ones will be discussed in this paper, which describes the thinking behind the ongoing compilation of the Lyon-Albuquerque Phonological Systems Database (LAPSyD) [15, 16] and draws some comparisons with other somewhat similar projects. such as PHOIBLE [20], SAPhon [19] and Segérer's African consonant inventory database [28]. The four issues are: How should a sample of languages be chosen? What range of data should be included? How to select between alternative descriptions? Should original (faithful to the source) or reinterpreted data be entered? These four issues are inter-related and decisions on one will often have implications for another.

2. HOW SHOULD A SAMPLE OF LANGUAGES BE CHOSEN?

The way this question is framed presupposes that the universe within which a sample will be selected is one of "languages", but of course this is a far from simple issue. There are no agreed criteria to decide whether two speech varieties differ sufficiently to be considered different languages. A common procedure in compiling cross-language samples is to restrict inclusion to no more than one member of a higher-level genetic grouping (e.g. [2, 13, 22, 23]). This procedure enables the language/dialect dilemma to be side-stepped, but undermines two potentially useful properties that a database might have, namely, first, reflecting the actual numbers of distinct languages within genetic groups so as to more closely represent the overall population of extant languages, and, secondly, being able to compare similarity within and between language families. It is often assumed, for example, that languages within the same family will be phonologically more similar to each other than to languages in other families (such similarity may even be part of the evidence for grouping languages into families). However, given the processes of language-internal change and convergence through contact, it is not necessarily the case that degrees of similarity follow genetic groupings. A more inclusive sampling provides the potential for testing the influence of family membership. Database users interested in examining a stratified sample can always construct an appropriate sub-sample from a larger set, but the larger picture cannot be seen from a pre-restricted selection. However, it still seems redundant to include speech varieties known to be mutually intelligible, such as the French of Paris and Québec, in part because of the costs in time and effort required to process and enter each sample.

There is also value in including major languages since users are likely to be interested in these, and in including languages that are found in other typological surveys so that, for example, possible interrelationships between characteristics of different types might be explored. Both these points were considered in the WALS project [3] and

influenced the selection of the core 100 and 200-language samples for that project. These languages as well as all those in the UPSID database [13] are included in LAPSyD (there is considerable overlap).

As in WALS, LAPSyD output is often visualized on a map; hence, other considerations also come into play. Maps can provide a useful guide to the language density of different parts of the world, so where there are large numbers of different languages close together — even if quite closely related, as in the case of the Bantu sub-group of Niger-Congo, or the Oceanic branch of Austronesian — it can be valuable to include a fair number of these so that this property emerges on the maps. Further, regions where there are few distinct languages should not appear unpopulated, so efforts to include languages in areas of low linguistic diversity, such as North Africa, are also valuable. In this connection it should be noted that language locations are identified in LAPSyD as points, as in WALS, not as areas, as in maps in the *Ethnologue* [12] or in the *World Language Mapping System* [33]. For living languages, these points generally represent a central point of the current distribution of speakers or the location of specific fieldwork, not locations prior to the major colonial expansions from the fifteenth century C.E. onwards (as in WALS). Major world languages are given a political home (Russian in Moscow, Mandarin Chinese in Beijing, Spanish in Madrid, English in London, etc).

3. WHAT RANGE OF DATA SHOULD BE INCLUDED?

The goal of including comparable data on a substantial number of languages imposes practical limitations on the range of data that can be covered. The greater part of an in-depth phonological study of an individual language is typically devoted to describing the alternations that occur in different contexts. These are often specific to given morphosyntactic environments and are difficult to generalize. Moreover, in-depth phonologies are available for a relatively small number of languages. It is a welcome development that many grammars of endangered languages have been published in recent decades, and that this kind of study is again a respectable enterprise for a linguistics Ph. D. However, many of these grammars devote only a handful of pages to phonological topics [14]. Most typically they provide a basic inventory of consonants and vowels, though sometimes only with very imprecise description of their pronunciation, plus some information on whether there is a tone or stress system and some comments on phonotactics. These basic data are therefore the data that are targeted for inclusion. However, LAPSyD differs from some broadly analogous projects, such as PHOIBLE [20] or SAPhon [19] in the inclusion of brief commentary fields on consonants, vowels, tone, stress, and syllable structure. This enables salient questions of interpretation or lacunae in the information to be flagged for the user.

4. HOW TO SELECT BETWEEN ALTERNATIVE DESCRIPTIONS?

For certain languages, there is essentially a single source of detailed published information. This is the case, for example, for Tukang Besi, Mosetén or Koromfe, where the grammars of Mark Donohoe [5], Jeanette Sakel [27] and John Rennison [26] respectively provide the only substantial works on these languages available at this time. But for others, there may be multiple sources which provide differing analyses among which a choice must be considered. These differences are likely to have three main sources.

First, although describing the same language, they may be describing different varieties, perhaps from different localities or different age-groups. That is, there are objectively different patterns in the phonology. Here the resolution is to decide on the targeted variety. For example, descriptions of standard French may include a contrast between low front and low back vowels, /a̟/ and /ɑ/ (e.g. Marchal [17]). However, many contemporary speakers do not make this contrast and have the same low central vowel /a/ in both the words “patte” and “pâte” which used to form a minimal pair. French as entered in LAPSyD reflects this newer pronunciation norm as reported *inter alia* by Landercy & Renard [11] or Fougeron & Smith [8].

Second, differences between descriptions may reflect differences in what might be called the scope of coverage. For example, many languages have sounds that occur only in a few exclamations or in onomatopeic or ideophonic words. Bentley & Kulemaka [1] note vowel length in ideophones in the Chichewa dialect of the Bantu language Nyanja, but Mchombo [18] makes no mention of vowel length. Here the choice might be dependent on how central such sounds seem to be to the language’s structure; that is, do they seem more like paralinguistic than linguistic features. A similar situation may exist with loanwords, which may include sounds or syllabic patterns that are not found in native vocabulary. One author may consider these to be integral to the language while another does not. For example, both Ham [9] and de Oliveira [4] agree

that in the Jê language Apinayé of Brazil /f/ only occurs in loans from Portuguese such as /famas/ "pharmacy". Ham does not include /f/ in her consonant chart, and does not discuss any variants in its pronunciation; de Oliviera does include /f/ and notes that, like other voiceless obstruents, it is allophonically voiced in weak positions. Is this sufficient to conclude that this sound is integrated into the phonology of Apinayé? The decision in this case was to exclude the segment from LAPSyD, based on two factors: first, the nature of the lexical forms cited seem far from central to Apinayé life; second — and even more subjective — an examination of a recording of the language supported Ham's analysis in several details, especially concerning the vowel system, and hence indirectly encouraged confidence in her analysis of the system.

Third, differences between descriptions may reflect varying analytical preferences or theoretical stances on the part of the authors. A simple case is represented again by Chichewa. Bentley & Kulemeka [1] list a total of 67 consonants, which includes 20 labialized ones, and 22 prenasalized ones. Mchombo [18] lists only 28, mostly since the labialized and prenasalized consonants of Bentley & Kulemeka are taken to be sequences of two separate consonants. It would be possible to say that both analyses are equally valid, or to design a database where the unit of entry is a language description, not a language. The latter is partly the approach of Ségerer [28] in his database of consonant inventories of African languages. For example, three listings of consonants drawn from different sources are given for Ibibio, showing 13, 15 or 16 segments. This variation is due to different theoretical models (generative/phonemic), different dialects, and different decisions on cases where contrast is limited. Such an approach shifts the responsibility of selecting an appropriate analysis for their purposes to users of the database, one they may not be well-equipped to bear.

In LAPSyD, an attempt is made to determine a single analysis which is to be preferred, as is also the case in the SAPhon database on South American languages [19]. SAPhon cites the recency of the description, the explicitness of support cited for the analysis, and the apparent linguistic sophistication of the describer as factors in selecting between analyses, but largely accepts a given analysis as it is presented by an author or contributor. Similar factors are also considered in LAPSyD, but the aim is further to harmonize descriptive principles and procedures as best as possible across all languages in the database. Because of the high value placed on this objective the analyses that appear in published sources on individual languages often seem to call for modification. This raises the important question of fidelity to sources.

5. SHOULD ORIGINAL (FAITHFUL TO THE SOURCE) OR REINTERPRETED DATA BE ENTERED?

Phonological databases differ in whether they aim, as the PHOIBLE database states (Moran et al [20]), to "[b]e faithful to the language description in the source document" or not. An approach emphasizing fidelity can easily result in inconsistencies. A simple case illustrates this. PHOIBLE lists the 'East Papuan' language Bilua as having the voiced stops /b, d, g/, based on the symbols in the consonant chart given in Obata [25]. Obata's text actually describes these stops as prenasalized except in initial position (the position where prenasalization is most often misheard and overlooked!), and the older spelling of the language name and the district where it is spoken was Mbilua. Hence LAPSyD lists these segments as /mb, nd, ŋg/. Another Solomon Islands East Papuan language, SavoSavo, is shown in PHOIBLE as having voiced prenasalized stops even though the ultimate source, Todd [29] used plain symbols such as /b, d/ for their transcription (cf also Wegener [31]). For this language the data in PHOIBLE was taken from UPSID, which had noted and transcribed the prenasalization verbally reported in Todd's text. Absence of plain voiced stops is an areal feature in languages of the Solomon Islands, but this becomes less apparent if Bilua and SavoSavo are treated differently due to being faithful to a data source rather than to the linguistic facts.

Zuni, an isolate spoken in New Mexico, provides an interesting case where the source description invites a re-analysis. The best-known reference on this language is Newman's short grammar from 1965 [21]. Newman proposes a phoneme inventory of 16 consonants and 10 vowels (5 long and 5 short). Word-medially a large number of consonant sequences occur, including identical clusters (geminates) and C+ʔ sequences, the latter realized as ejectives when C is a voiceless stop or affricate. In Newman's view, four of these C+ʔ sequences also occur word-initially. But when this distributional restriction is taken into account an interpretation of this set as unitary ejectives, as hinted at by Walker [30] and de facto adopted by Nichols [24], seems preferable. Furthermore, if /kʼ kʷʼ tsʼ tʃʼ/ are units this simplifies syllable structure, since no CC- onsets occur, and no word-medial -CCC- strings occur

under this analysis. Due also to the absence of codas after long vowels — a fact which is obscured if ejectives are analyzed as biphonemic and hence appear to be heterosyllabic — Newman's proposed maximal syllable structure of CCV:CC is reduced to CV{:, C}. Hence LAPSyD shows this as the syllable canon and has an inventory of 20 consonants for Zuni including the 4 ejectives (as also earlier in UPSID). Again, this clarifies an areal pattern among languages of the American Southwest which predominantly include ejectives in their inventory.

At the time Newman was writing linguists were often interested in reducing the contrastive inventory to the minimum. So it is perhaps surprising that he did not reduce the vowel inventory to 5 plus a "phoneme" of length as Granberry [10] and Walker did [30]. In UPSID when all the short vowels of a language had a long counterpart long vowels were not entered as separate phonemes (since the focus there was on the number of distinct vowel *qualities*). PHOIBLE, taking data from UPSID, thus lists Zuni with 5 vowels and 20 consonants. But it also lists the segments reported in the Stanford Phonology Archive for Zuni, which includes all 10 vowels but also lists as separate segments all the geminate consonants and the allophones of the voiceless stops and affricates which Newman heard as aspirated, yielding an inventory of 44 consonants. Both of these descriptions of Zuni are faithful to the immediate sources, but the sources in this case are both secondary, and vary from the original source description in different ways. No guidance is provided as to which of these might best be taken as a basis for comparing Zuni to other languages.

Among the databases considered here, Ségerer's compilation on consonant systems of African languages is the most literally faithful to the sources. A typical entry includes a scanned page from a primary source showing a labeled consonant chart. Beside this a 'harmonized' chart is shown with numbered rows and columns and a more standardized IPA-based transcription replacing ad hoc or idiosyncratic symbols in the original, but without any further re-interpretation. As noted earlier, this database, like PHOIBLE, often includes multiple descriptions of the same language. Thus, the Atlantic language Basari is represented by two entries, based on two publications by Ferry [6, 7]. One shows an inventory of 31 consonants, the other 36. Between 1961 and 1968 Ferry changed her mind about the analysis of labialized velars, preferring in the later publication to interpret them as unitary elements rather than /Cw/ sequences, a view supported in more recent work by Winters & Winters [32]. There are no other sequences with glides and no secondary articulations occur at other places of articulation. These points are among classic arguments for segmental unity, and hence the larger consonant inventory is the preferable alternative (as in the Zuni case). The two analyses cited by Ségerer are both faithful to their sources but not equally informative about the language.

There are many cases beyond these simple examples where it seems that critical judgment needs to be applied to data in a source, and the database compiler is in the best position to apply this judgment. All entries in LAPSyD are the result of careful reading of the sources, and where it seems appropriate a reanalysis or reinterpretation of the data. This procedure assuredly provides a much more secure basis for cross-language comparison.

6. CONCLUSIONS

In this paper some of the design considerations that have guided the ongoing construction of the LAPSyD database have been briefly presented. LAPSyD aims to include a wide range of languages from which subsamples can be drawn when desirable. It limits the data to basic information of the kinds broadly available in any general language description. It selects a single 'preferred' analysis for each language, and aims to harmonize descriptive approaches across all the languages so as to provide a sound basis for cross-language comparisons. It is hoped that these features will make this a useful resource for typological linguistic research, as well as a valuable instructional resource.

In order to facilitate these functions, LAPSyD also aims to provide a comprehensive suite of tools to search the contents of the database and to select and export the desired information. As these tools are refined, interactive interrogation of a large body of knowledge will beome possible.

7. REFERENCES

[1] Bentley, M. & A. Kulemeka. 2001. *Chichewa*. Lincom Europa, München and Newcastle.

[2] Bybee, J., R. Perkins & W. Pagliuca. 1994. *The Evolution of Grammar*. University of Chicago Press, Chicago.

[3] Comrie, B., M. S. Dryer, D. Gil, M. Haspelmath. 2013. Introduction. In Dryer, M. S. & M. Haspelmath (eds.) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (http://wals.info/chapter/s1, Accessed 2015-01-30.)

[4] de Oliveira, C. C. 2005. *The language of the*

Apinajé people of Central Brazil. Ph. D. dissertation, University of Oregon. Available at http://etnolinguistica.wdfiles.com/local--files/tese:oliveira-2005/oliveira_2005.pdf
[5] Donohoe, M. 1999. *A Grammar of Tukang Besi*. Mouton de Gruyter, Berlin.
[6] Ferry, M-P. 1961. Le basari. In Perrot, Jean (ed.), *Les langues dans le monde ancien et moderne, volume 1*. Centre National de la Recherche Scientifique, Paris: 55-63.
[7] Ferry, M-P. 1968. Deux langues Tenda du Sénégal Oriental: Basari et Bedik. *Bulletin de la Société d'Études Linguistiques et Anthropologiques de France* 7: 1-62.
[8] Fougeron, C. and C. L. Smith. 1999. French. *Handbook of the International Phonetic Association.* Cambridge University Press, Cambridge: 78-81..
[9] Ham, P. 2009 [1961]. Apinayé phonemic statement: preliminary version. Associação Internacional de Linguística-SIL Brasil, Anápolis, Brazil. Available at http://www.sil.org/americas/brasil/publcns/ling/AYPhonem.pdf
[10] Granberry, J. 1967. *Zuni Syntax*. Ph. D. dissertation, State University of New York, Buffalo.
[11] Landercy, A., & R. Renard. 1977. *Éléments de Phonétique*. Didier, Bruxelles.
[12] Lewis, M. P., G. F. Simons, & C. D. Fennig (eds.). 2015. *Ethnologue: Languages of the World, 17th edition*. Dallas, Texas: SIL International. (Online at: http://www.ethnologue.com. Accessed 2015-04-27.)
[13] Maddieson, I. 1984. *Patterns of Sounds*. Cambridge University Press, Cambridge.
[14] Maddieson, I. 2003. Field phonetics. In J. Larson & M. Paster, eds., *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society,* 411-429.
[15] Maddieson I., S. Flavier, E. Marsico & F. Pellegrino. 2015. *LAPSyD: Lyon-Albuquerque Phonological Systems Database*. http://www.lapsyd.ddl.ish-lyon.cnrs.fr/lapsyd/
[16] Maddieson, I., S. Flavier, E. Marsico, C. Coupé, & F. Pellegrino. 2013. LAPSyd: Lyon-Albuquerque phonological systems database. *Proc. Interspeech 2013*, Lyon: 3022-3026.
[17] Marchal, A. 2007. *La Production de la Parole.* Lavoisier, Paris.
[18] Mchombo, S. 2005. *The Syntax of Chichewa*. Cambridge University Press, Cambridge.
[19] Michael, L., T. Stark, and W. Chang (compilers). 2012. South American Phonological Inventory Database v1.1.3. Survey of California and Other Indian Languages Digital Resource. Berkeley: University of California. (Online http://linguistics.berkeley.edu/~saphon/en/ Accessed on 2015-02-01.)
[20] Moran, S., D. McCloy, R. Wright (eds.) 2014. PHOIBLE Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Online http://phoible.org, Accessed on 2015-02-02.)
[21] Newman, S. 1965. *Zuni Grammar*. University of New Mexico Press, Albuquerque.
[22] Nichols, J. 1992. *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago.
[23] Nichols, J. 1998. The origin and dispersal of languages: Linguistic evidence. In *The Origin and Diversification of Language*, ed. N. G. Jablonski & L. C. Aiello. California Academy of Sciences, San Francisco :127-170.
[24] Nichols, L. 1997. *Topics in Zuni Syntax*. Ph. D. dissertation. Harvard University.
[25] Obata, K. 2003. *A Grammar of Bilua: A Papuan Language of the Solomon Islands (Pacific Linguistics 540)*. Research School of Pacific and Asian Studies, Australian National University, Canberra.
[26] Rennison, J. R. 1997. *Koromfe*. Routledge, London and New York.
[27] Sakel, J. *A Grammar of Mosetén*. Mouton de Gruyter, Berlin.
[28] Ségerer, G. Inventaires consonantiques des langues africaines. (Available online at http://www.guillaumesegerer.fr/Ling/Consonnes/ Accessed on 2015-02-01.)
[29] Todd, E. M. 1975. The Solomon language family. In S.A. Wurm (ed.), Papuan Languages and the New Guinea Linguistic Scene. New Guinea Area Languages and Language Study, Vol. 1. Pacific Linguistics, Series C, No. 38. Australian National University, Canberra.
[30] Walker, W. 1972. Toward the sound pattern of Zuni. *International Journal of American Linguistics* 38: 240-259.
[31] Wegener, C. 2009. A grammar of SavoSavo: A Papuan language of the Solomon Islands. Ph. D. dissertation. Radboud University, Nijmegen. Revised edition 2012, published by De Gruyter Mouton, Berlin.
[32] Winters, James & Patricia Winters. 2002. *Onëyanoŋ: Une Etude Phonologique du Bassari*. Société Internationale de Linguistique, Dakar.
[33] World Language Mapping System. Global Mapping International, Colorado Springs.