LAPSyD: Lyon-Albuquerque Phonological Systems Database

Ian Maddieson[1], *Sébastien Flavier*[2], *Egidio Marsico*[2], *Christophe Coupé*[2], *François Pellegrino*[2]

[1] Department of Linguistics, University of New Mexico, Albuquerque, NM, USA
[2] Laboratoire Dynamique Du Langage, CNRS – Université de Lyon, Lyon, France

ianm@berkeley.edu, {Sebastien.Flavier, Christophe.Coupe, Egidio.Marsico}@ish-lyon.cnrs.fr, Francois.Pellegrino@univ-lyon2.fr

Abstract

LAPSyD, the Lyon-Albuquerque Phonological Systems Database, is an online phonological database equipped with powerful query, mapping and visualization tools. It stems from the UPSID and WALS databases, enhanced with newly validated data not only covering segmental inventories but also syllable structures, stress and tonal systems. In its current version it covers around 700 languages and it is accessible at http://www.lapsyd.ddl.ish-lyon.cnrs.fr. This paper provides a description of the data structure in LAPSyD and the features of the interface. Brief illustrations of the types of analysis that can be done with this tool are provided, exploiting the ability to cross-reference data on segments, other phonological properties and language location.

Index Terms: phonological inventory, phonological typology, tonal systems, syllable type, UPSID, acoustic adaptation hypothesis, multilingual database

1. Introduction

The quality of any typological investigation strongly depends on the database that supports it; thus, the design of such a database is important. Phonological typology is no exception. The pioneering work of Trubetzkoy [1] used a “memory based” database of around 30 languages. Since then improvements have been made in all ways, in number of languages, contents, architecture, accessibility and scope of research. Following the path opened by the Stanford Phonology Archive project [2, 3], much 20th-century phonological typology focused on the identification of articulatory and/or perceptual constraints explaining the structure of phonological inventories. Much of this research was based on the UPSID database [4, 5, 6], for a long time the largest balanced sample of phonological inventories available, covering 451 languages. Of course, phonological typology is also concerned with matters beyond segment inventories, and extends to many other aspects including syllable structure, suprasegmentals, and phonological processes. The Lyon-Albuquerque Phonological Systems Database (LAPSyD) now being developed extends the scope of data covered in UPSID so that certain of these other matters are also included, specifically, information on tone, stress and syllable structure.

In comparison to most other currently available cross-language phonological databases this represents a richer range of data on each language. Table 1 lists a number of these other resources and highlights how LAPSyD exceeds their scope.

The next section consists in a presentation of the database features (languages, linguistic data and, user modules). Section 3 presents several analyses performed to illustrate the richness of the database.

2. LAPSyD Description

2.1. Basic Design

Although LAPSyD is a direct extension of UPSID, the phonological information in LAPSyD is more extensive and has been extensively updated based on more recent research. In addition to inventories of consonants and vowels, an outline of the permitted syllable structures and brief comments on the role played by tone and stress (accent) is included, together with a classification of the complexity of the syllable canon and the tone system, and on the role of stress. Commentary fields on all these aspects briefly discuss the interpretations made and remaining ambiguities. The database includes a geographical module providing mapping of either the entire sample or of query results. Unlike UPSID, LAPSyD is not designed to be a balanced sample: some very closely related languages are included, some of the better-studied families are over-represented, and languages from less well-known areas are necessarily under-represented. However, a representative sample can be constructed by making a selection among the languages included; for example, all languages in LAPSyD which occurred in UPSID are flagged.

Table 1. *Non exhaustive list of current existing phonological databases, white names indicate online ones.*

	Languages			**Phonological content**		**Database Features**	
Name	Number	Area Specific / Worldwide	Genetically Balanced	Segments / Syllables / Tones	Phonetic Feature system	Map Visualization Module	Query Module Simple/Complex
SALA [7, 8]	200	Specific	no ?	**yes** / no / no	no	no	**yes** / no
SAPhon [9]	359	Specific	no	**yes** / no / no	no	no	**yes** / no
PHOIBLE [10]	~1000	Worldwide	no	**yes** / no / no	no	no	**yes** / no
P-BASE [11]	>500	Worldwide	no	**yes** / no / no	no	no	**yes** / no
XTone [12]	80	Worldwide	no	no / no / **yes**	no	no	**yes** / no
PhonQuery [13]	22	Worldwide	not yet	**yes / yes / yes**	no	no	**yes / yes**
UPSID [4, 5]	451	Worldwide	**yes**	**yes** / no / no	yes	no	**yes** / no
ULSID [14]	22	Worldwide	no	no / **yes** / no	no	no	**yes** / no
LAPSyD	~700	Worldwide	**yes** / no	**yes / yes / yes**	**yes**	**yes**	**yes / yes**

LAPSyD is being made available for general interest and as a research tool. It is planned to continue to expand it both by adding further languages and by increasing the richness of information about each individual language included.

2.2. Language selection

The primary criterion for inclusion of a language in LAPSyD is the availability of a reliable description of its main phonological characteristics based on first-hand experience with the language and prepared by someone with an adequate level of linguistic sophistication. Languages no longer currently spoken may be included if the documentation obtained before their extinction is adequate to make a reasonably reliable basic phonological analysis.

Both geographical and genetic factors play a role in selection of languages. Virtually all the adequately-described languages spoken in large areas which are sparsely-populated or have little linguistic diversity (e.g. North Africa, Siberia) are likely to be selected in order to fill the space on maps. Where language density is greater, language selection is influenced by impressions of language diversity. For example, the Bantu zone of east, central and southern Africa is less densely sampled than New Guinea, as there is greater genetic diversity of languages in the latter area.

The information on each language comes primarily from published or publicly-available technical linguistic literature, such as grammars, dissertations, journal articles and dictionaries. Detailed bibliographical references are given for the items relied on for each language. In a few cases, data is based on or supplemented by personal fieldwork by the compiler, personal communications from others or resources that may be accessible on the web.

2.3. Data

The primary goal of the database is to represent the segmental and prosodic contrasts that form lexical distinctions in each language, together with basic information on phonotactics. The phonological description of each language is reviewed to standardize the analyses as far as possible. This is done in order to remove differences that have to do with choice of theoretical model, transcriptional preferences and other issues that create apparent rather than real differences between the languages included, or disguise real differences that do exist. This homogenization of the data is regarded as an important and valuable feature of the database. Obviously, the data in LAPSyD is subject to limitations on the information available in the sources consulted. It is quite common, for example, to read a description that mentions long vowels or nasalized vowels as contrastive but which fails to state how many such vowels exist, or to find no explicit statement on syllabic structure. By examining words cited as examples or studying a lexicon it may be possible to construct an idea of the syllable canon, but not all such lacunae can be filled.

All of the vowel and consonant segments referenced in the database are given a unique IPA transcription and featural description. This description in features enables searches to be conducted for all occurrences of segments with individual features or sets of features and to look for co-occurrences or patterns of complementary distribution and other properties of the inventories at the featural level. The features used to define the segments catalogued in LAPSyD are based on traditional phonetic terminology, such as that embodied in the charts of the International Phonetic Alphabet (IPA). Within the limitations allowed by the source descriptions, LAPSyD aims to represent all the within-language contrasts encountered in each language with as much fidelity to cross-language comparison as possible.

Each language is located at a particular location (as in the *World Atlas of Language Structures* [15, 16]) and is also assigned to one of six major areal-genetic groupings so that analyses can be repeated within each group ([17, 18]. Table 2 lists the most salient data contained in LAPSyD sorted by type in the right column, with metadata in the left column.

Table 2. *Data in LAPSyD.*

Data on the language	Phonological data
Name and alternates	Vowel inventory
IS0-639 code	Consonant inventory
Classification	Syllable structure
Localization	Tone system
Areal / genetic grouping	Stress (accent) system
Sources consulted	Frequency data
Links to recordings, etc.	

2.4. Architecture, Interface and modules

LAPSyD is a web application using a dynamic scripting language such as PHP with Ajax technology. The web application is based upon a relational database management system (MySQL) to store data, and some third-party libraries, such as Mootools, jQuery-sparkline and Polymaps, all Javascript Libraries. The server hosting LAPSyD is in a Windows environment using Apache software.

The only requirement to work with LAPSyD is to use an HTML5 browser with CSS3 support such as Firefox, Chrome (compatibility checks are made as a priority for these two browsers) or Safari. Consonant and vowel charts are displayed in a format similar to that adopted for the IPA. To view these charts correctly the installation of a UNICODE font such as Charis SIL is highly recommended.

LAPSyD is structured in a series of different modules allowing a user to simply browse through the data (the ‘Data’ menu) or to make simple or complex queries (the ‘Query’ menu). In addition, a module not open to the public allows an Administrator to modify existing entries or to add new data (the ‘Management’ menu). There are three different user levels: User, Supervisor and Administrator. Any member of the public can be a User.

A particularly interesting feature of LAPSyD is the distinction between three layers of data. ‘Public’ data are those available for any user; ‘Private’ data is information submitted to enrich the database which has not yet been validated by a LAPSyD administrator and rendered public; and finally, there is ‘Hidden’ data, which allows a user to manage data that he or she has entered for their personal use. Using this feature, a user can add data on additional languages in a format compatible with the LAPSyD format that only they will be able to see. This last feature means that LAPSyD can be used as a personal phonological database manager.

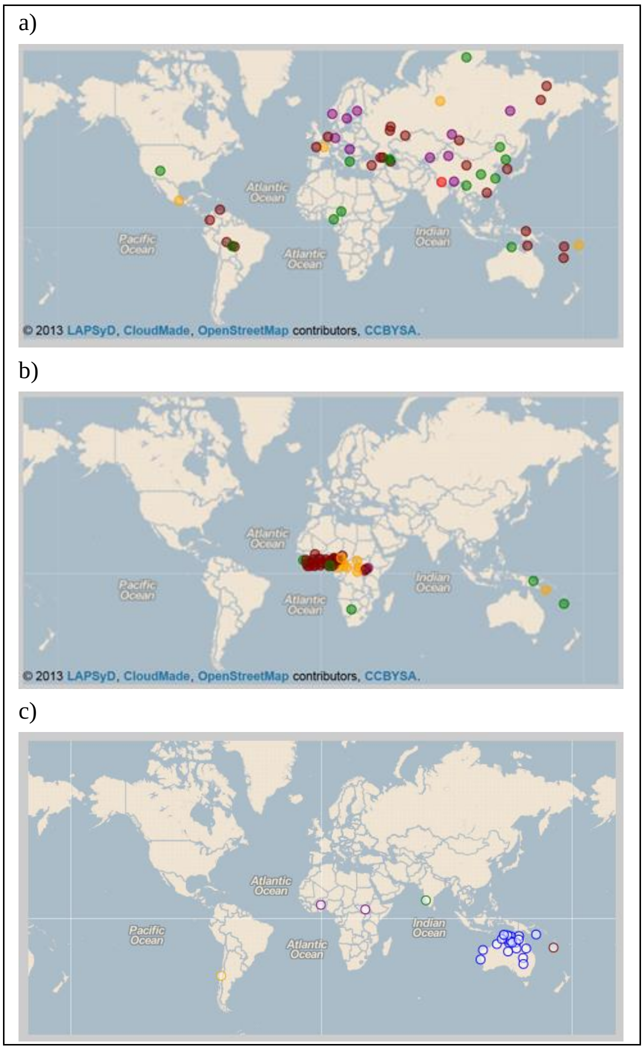
Queries to the database can be constructed based on individual segments (by clicking on an IPA symbol), or based on features and intersections of features, or on the syllabic, tonal and stress data, as well as on intersections of these and of these with segmental properties. Results of the queries can be exported as Excel files or visualized in various graphical ways, such as on maps or as histograms or contingency tables.

3. Phonological typology with LAPSyD

3.1. New approaches to old issues

It is well-known that certain types of segments have markedly restricted geographical distributions. LAPSyD offers the possibility to easily visualize such distributions on a map. Figures 1a to 1c illustrates some well-known distributions.

Figure 1: *Well-known areal distributions of segments. a) front rounded vowels, b) labial-velar stops and c) rich place contrast for systems of nasals*



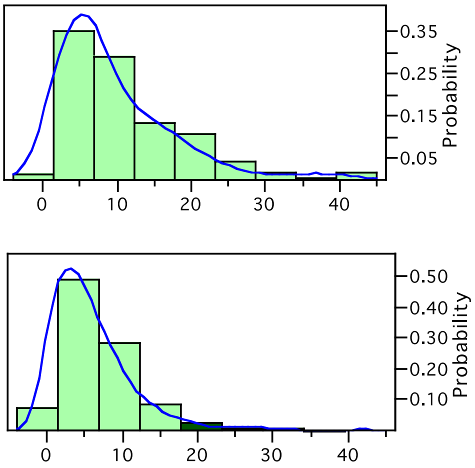
Front vowels are most often unrounded, but front rounded vowels occur in some languages (they are present in 54 of the 692 languages sampled). As figure 1a shows, front rounded vowels are mainly found in the old world, from East Asia to Europe. This led Crothers [19] to think that this situation resulted from contact-induced influence of Altaic on European languages. Although this is possible, the current distribution suggests that several independent appearances of front rounded vowels are to be envisaged.

Figures 1b and 1c illustrate two particularly restricted distributions, namely, the labial-velar stops /kp/ and /gb/, mostly confined to equatorial Africa, and extensive series of nasals contrasting in place of articulation, predominantly found in Australian languages. Labial-velar stops are found in languages of several different families in Africa, suggesting an areal diffusion, whereas the Australian proliferation of nasals may be an inherited trait.

3.2. Relation between segmental and syllabic complexity

Queries to the LAPSyD data can also be based on custom combinations of features defining desired sets. A feature combination defining the Complex (including Elaborated) consonant class as defined by [20] was constructed, and the occurrence of consonants so defined was examined in relation to syllable structure. Results suggest that there is a significantly higher probability that languages with Complex syllable structure have more consonants belonging to the Complex and Elaborated classes. Languages with Complex syllable structure have a mean of 9.63 of the consonants so specified, whereas languages with Simpler syllable structure have a mean of 5.38 (6.23 for Moderate languages, 4.75 for Simple languages, but this difference is not significant by Tukey-Kramer HSD, so these classes have been merged). One way of visualizing the probability distribution is by comparing the two histograms in Figure 2. The first shows Complex syllable structure languages (n = 212); the second, languages with simpler maximum syllable complexity (n = 435).

Figure 2: *Number of Complex or Elaborated consonants in languages with Complex syllable structure (upper panel) and Simpler syllable structure (lower panel).*



3.3. Relation between complex nuclei and coda complexity

LAPSyD also provides a good tool to test contingency relationships. Queries can be constructed to examine how properties of various types intersect and the results output as a contingency table. We tested, and were able to reject, a hypothesis relating the presence of long vowels in a language’s inventory to the presence of more complex syllable codas. It seems a reasonable expectation that languages might balance a tendency to have heavy rhymes based on vowel length against having syllable weight due to complex codas, defined as those allowing two or more consonants in the coda position. In fact, as the data in Table 3 shows, long vowels occur in about 33% of the languages, regardless of whether they have only simple codas or complex ones.

Table 3: *Contingency table of long vowels/simple codas*

	long vowels	no long vowels	row total
no/simple coda	**135 (33%)**	**95 (33%)**	*230*
complex coda	**269 (67%)**	**183 (67%)**	*462*
column total	*404*	*288*	*692*

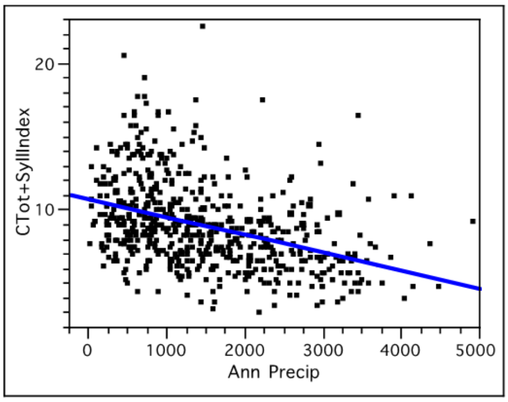
3.4. Testing the Acoustic Adaptation Hypothesis

A further type of analysis that can be done using data from LAPSyD concerns cross-referencing linguistic properties with non-linguistic ones using the language location coordinates to link with databases containing, say, climatic and ecological information. Ideas suggested by the Acoustic Adaptation Hypothesis [21, 22, 23] as applied to language [e.g. 24, 25] have been explored, drawing on data available from several datasets, such as elevation and rugosity ([26], distance to water [27], length of the growing period of plants [28], yearly mean temperature and total precipitation [29] and tree cover [30]. Values for these different variables were taken at the point locations of each language available in LAPSyD (A gross approximation for languages spoken over large areas, but generally valid for “smaller” languages, as ecological conditions are relatively similar for all speakers).

The basic idea of the AAH as applied to human spoken languages is that, among other factors, the physical environment in which a language is (or was) habitually used has an influence on its phonetic and phonological characteristics because of environmental effects on the transmission of the signal. A priori, a number of factors such as higher heat or humidity, greater coverage of vegetation, or stronger prevailing winds, might influence language structure in that each selectively impedes faithful transmission of higher frequency components, more typical of consonants than of vowels. A number of correlations have been explored: we report one of the most salient. An index combining consonant inventory size and syllable structure complexity was constructed to reflect how “consonant-heavy” each language is, in terms of number of consonant contrasts and how these are deployed in forming simple or complex syllables. This index combines a syllable complexity index ranging from from 1 to 8 added to the consonant inventory size divided by 4 (in order to roughly equalize the influence of the two factors). Annual precipitation is a significant predictor of this index (as also are mean annual temperature, and rugosity, but less strongly).

Figure 3 plots the fit for 569 languages for which values were available (r^2 = .156, p < .0001). Broadly, the more it rains the more likely a language is to have fewer consonants and simpler syllable structures. Rainfall may here be partially a proxy for natural dense vegetation cover and high humidity.

Figure 3: *Linear fit between annual precipitation and consonant-heaviness (3 outlier cases not plotted).*



LAPSyD also allows for exploration of the generality of an effect through subsetting by area or language family. In this case in the 6 macro-areas into which languages are grouped, the same direction of correlation seen in Figure 3 is found in 5 and in 4 it is significant at p < .05 or better. The exception is North America which shows an opposite correlation. In the 11 language families which are represented by at least 10 languages the correlation is found in 6, although it only attains statistical significance in 3. No significant counterexamples are found. The effect therefore cannot be easily dismissed as an artifact of areal or genetic differences between languages

4. Conclusions

This paper has presented a brief outline of the LAPSyD database and selected sample analyses that exploit the richness of the date included. These analyses have illustrated the power of this system to combine information from the segmental, featural, syllabic and geographic levels of the data to provide insights that go beyond simple information on consonant and vowel inventories.

5. Acknowledgements

We gratefully acknowledge the support of various institutions that have contributed to the development of this project (including Collegium de Lyon and LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon, for its financial support within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French government operated by the National Research Agency (ANR).

6. References

[1] Trubetzkoy, N. Grundzüge der Phonologie. TCLP VII. Prague, 1939.

[2] Sherman D. and Vihman, M. "The Language Universals Phonological Archiving Project: 1971-1972," Working Papers in Language Universals, No. 9. 163-17, 1972

[3] Greenberg, J.H. et al, [Ed]. Universals of Human Language, Vol. 2, Phonology. Stanford: Stanford University Press, 1978.

[4] Maddieson, I. Pattern of Sounds. Cambridge, UK: Cambridge University Press, 1984

[5] Maddieson, I., & Precoda, K. Updating UPSID. UCLA Working Papers in Phonetics 74: 104–111. Department of Linguistics, UCLA, 1990.

[6] Reetz, H. Web interface to UPSID. http://web.phonetik.uni-frankfurt.de/upsid_info.html, 1999.

[7] Hartell, R. L. Alphabets des langues africaines. UNESCO. SIL, 1993.

[8] Chanard, Christian. Systèmes Alphabétiques des langues africaines. Online: http://sumale.vjf.cnrs.fr/phono/, 2006.

[9] Michael, Lev, Tammy Stark, and Will Chang (compilers). South American Phonological Inventory Database v1.1.2. Survey of California and Other Indian Languages Digital Resource. Berkeley: University of California. http://linguistics.berkeley.edu/~saphon/en/, 2012.

[10] Moran, Steven. 2012. Phonetics Information Base and Lexicon. PhD thesis. University of Washington. http://phoible.org/

[11] Mielke, Jeff. The Emergence of Distinctive Features. Oxford: Oxford University Press, 2008.

[12] Hyman, Larry M. 2001. "Tone systems". In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, & Wolfgang Raible (eds), Language typology and language universals: An international Handbook, vol. 2, 1367-1380. Berlin & New York: Walter de Gruyter. http://xtone.linguistics.berkeley.edu/index.php

[13] Kehrein, W. & Knaus, J. http://www.online.uni-marburg.de/phonquery/index.php

[14] Vallée, N., Rossato, S., & Rousset, I. Favoured syllabic patterns in the world's languages and sensorimotor constraints. In Pellegrino, F., Marsico, E., Chitoran, I., Coupé, C. (Eds.). Approaches to Phonological Complexity (pp. 11-39). Berlin: Mouton de Gruyter , 2009.

[15] Haspelmath, M., Dryer, M. S., Gil, D. & Comrie, B. [Ed], World Atlas of Language Structures, Oxford and New York: Oxford University Press, 2005.

[16] Dryer, M. S. & Haspelmath, M. The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. Available online at http://wals.info/, 2011.

[17] Dryer, M. S. "Large linguistic areas and language sampling". Studies in Language 13: 257-292, 1989.

[18] Maddieson, I. "Correlating phonological complexity: data and validation." *Linguistic Typology* 10: 108-125, 2006.

[19] Crothers, J. Typology and universals of vowel systems. In [3]: 93-152, 1978.

[20] Lindblom, B. & Maddieson, I. Phonetic universals in consonant systems. In C. Li & L. M. Hyman [Ed], Language, Speech and Mind. Routledge, London. 62-78, 1978.

[21] Morton, E. S. "Ecological sources of selection on
avian sounds." *American Naturalist* 109, 17–34, 1975.

[22] Wiley, R. H. & Richards, D. G. "Physical constraints on acoustic communication in the atmosphere: Implications for the evolution of animal vocalizations." Behavioral Ecology and Sociobiology 3: 68-94, 1978.

[23] Ey, E., & Fischer, J. "The 'Acoustic Adaptation Hypothesis' - A review of the evidence from birds, anurans and mammals." Bioacoustics 19: 21–48, 2009.

[24] Fought, J.G., Munroe, R.L., Fought, C.R., Good, E.M. (2004). Sonority and climate in a world sample of languages: Findings And Prospects. Cross-Cultural Research 38: 27-51, 2004.

[25] Munroe, R.L., Fought J.G., Macaulay R.K.S. Warm climates and sonority classes: Not simply more vowels and fewer consonants. Cross-Cultural Research 43: 123-133, 2009.

[26] Amante, C., & Eakins, B. W. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24, 2009.

[27] Kummu, M., De Moel, H., Ward, P. J., & Varis, O. How close do we live to water? A global analysis of population distance to freshwater bodies. PloS One, 6(6), e20578. doi:10.1371/journal.pone.0020578, 2011.

[28] Van Velthuizen, H., Huddleston, B., Fischer, G., Salvatore, M., Ataman, E., Nachtergaele, F. O., Zanetti, M., et al. Length of growing period zones 1901-1996. In Mapping biophysical factors that influence agricultural production and rural vulnerability (pp. 7–8). Rome: Food and Agriculture Organization (FAO) & International Institute for Applied Analysis (IIASA), 2007.

[29] Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25(15): 1965–1978.

[30] The International Steering Committee for Global Mapping (ISCGM)/Geospatial Information Authority of Japan, Chiba University and collaborating Organizations. Vegetation (Percent Tree Cover). Global Map V.1 (Global version). Retrieved from http://www.iscgm.org/cgi-bin/fswiki/wiki.cgi, 2008.