LAPSyD

Lyon-Albuquerque Phonological Systems Database

Contents

Contents

1. **Introduction.**

This database contains searchable basic information about the phonological systems of a substantial number of different languages around the world. It aims to include languages from different geographic areas and language families in approximate proportion to their density. In this, and a number of other aspects LAPSyD draws on the experience of compiling the World Atlas of Linguistic Structures (WALS) (Haspelmath et al 2005, Dryer & Haspelmath 2013). Unlike its predecessor UPSID, described in Maddieson (1984) and other works, 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, appropriately structured representative samples can be constructed by making a selection among the languages included and tools for doing so form part of the LAPSyD framework. Languages in LAPSyD which occurred in UPSID or appear in WALS are identified in saved sample lists.

The basic unit of entry is a language. For each language in the database, a list of the contrastive vowel and consonant segments is given, together with some commentary on questions of interpretation in the 'consonant notes' and 'vowel notes' fields. In addition, a descriptive outline of the permitted syllable structures and brief comments on the role played by tone and stress (accent). Categorical labels for syllable structure, tone and accent are also provided, which are explained in more detail in ¶ 3.5 3.6 3.7. A count of the number of consonants, the total number of vowel nuclei, and of the number of basic vowel qualities is also provided. Each segment is assigned a description in terms of phonetic features, enabling the database to be searched for classes of segments. The features used are outlined in ¶ 4. Familiarization with this feature set is necessary to understand how contrasts are encoded in LAPSyD and the limits that the feature set imposes on their representation.

The languages are identified by a primary name and by an ISO-639 three-letter code wherever possible. In some cases alternate names are also given when these are in common use. The area where the language is spoken is described in a text field and the language is assigned a point location specified by latitude and longitude coordinates. An abbreviated classification of the language by family is provided and languages are also assigned to one of six major geographic/genetic groupings. Most of the data is drawn from published sources

or publicly available documents, such as dissertations, and the sources on which the description is based are fully cited. This 'metadata' on the languages is described in more detail in ¶ 2.

It is emphasized that each language included is represented in LAPSyD by a 'snapshot' of how it was spoken at a particular time and place by particular individuals or groups. Any language is an ensemble of varieties in constant evolution, and a different choice of time or place of observation could yield a very different picture of the phonology of the language. Morover, the data as presented in the sources is filtered by the compiler of this database in an effort to provide a uniform style of analysis, particularly as it relates to the inventories of consonants and vowels. This homogenization of the data is regarded as an important and valuable feature of the database. Matters of interpretation are described in more detail in ¶ 3.

LAPSyD is 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. **Language selection and metadata**

2.1. Criteria for language inclusion.

The primary criterion for inclusion of a language in LAPSyD is the availability of what appears to be 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. The most satisfactory sources are those that provide a phonetically-informed description of the pronunciation and provide explicit arguments for segmentation and contrast and an explicit discussion of syllable patterns and prosodic contrasts. Disappointingly few language descriptions meet these simple criteria. In order to include a wider range of languages, less satisfactory sources must often be used.

Languages no longer currently spoken may also be included if the documentation obtained before their extinction is adequate to make a reasonably reliable basic phonological analysis. Languages from areas affected by particularly severe language loss before the 20th century, such as the east coast of North America and much of the south and east of Australia, but for which adequate data exists for 'salvage' descriptions to be made, may be included with a lower threshold of reliability. Examples include Timucua (tjm) from Florida, USA, and Biri (bzr) from Queensland, Australia.

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 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.

No two varieties that are considered to belong to the same language are included; however, this criterion is imprecise as no clear distinction can be drawn between a difference of language and a difference of dialect. For example, only one variety of English (eng) is included despite very considerable phonological differences, especially within the British Isles, but both Moroccan (ary) and Egyptian Arabic (arz) are included based on claims of mutual unintelligibility of colloquial speech styles.

2.2. Language names.

Each language is indentified by a single primary name. When there is an established English name for the language this is used, e.g. German (deu), Basque (eus), Assamese (asm), Navajo (nav). In most other cases a conventionalized spelling of an indigenous name is usually preferred, following the general trend in most recent scholarly work. This may result in familiar names not being used as primary identifiers of a language: Lappish is now known as Saami (sma), Cambodian as Khmer (khm), Yurak as Nenets (yrk), Nootka as Nuuchahnutlh (noo). When a language is newly endowed with an orthography, subsequent literature frequently employs the name and spelling sanctioned by the authority establishing the orthography. Hence !Xóõ becomes !Khoon (nmn) and Yeletnye becomes Yélî Dnye (yle). Choice of the name to use is thus a judgment weighing familiarity, 'correctness' and guesses about how the language is likely to be referred to in the future.

The primary name may be followed by a modifier, separated by a comma, specifying a particular dialect or variety. This is particularly the case when the familiar name covers varieties that are different enough that they are clearly separate languages (or might be considered so). For example, Nahuatl and Chinantec are both represented by more than one variety which are distinguished by a modifier following the primary name. Examples are Nahuatl, North Puebla (ncj) and Chinantec, Lealao (cle).

On the other hand, a modifier preceding a name is a full part of the language name. For example West Makian (mqs) is not a western variety of a Makian language but a language that is spoken on the western half of the island of Makian (and some nearby islands) in the Maluku province of Indonesia.

A language name may also be followed by a disambiguator enclosed in parentheses. This is most often used to clarify which language is meant when more than one language can be referred to by the same spelling of a name. Such homographs are usually disambiguated by adding the name of the country. For example, Ika (Colombia) distinguishes the Chibchan language of that name spoken in Colombia (arh) from the Igboid language Ika spoken in Nigeria (ikk).

Alternative names are supplied for many languages when these are reasonably familiar, especially when they have been used in the linguistic literature. For example, earlier publications on the Waorani (auc) language of Ecuador referred to it as Auca. This is actually a derogatory exonym, but it is useful to be able to connect this name with the preferred alternative in view of its use in earlier literature. For the most part, however, variants of a name that are essentially just alternative spellings are not listed.

2.3. Language codes.

Each language, apart from a handful of exceptions, is also identified by its ISO-639 code. These codes are lower-case strings of three letters, often based on the name of the language (but sometimes arbitrary). These codes are a useful way to identify whether different sources and databases are referring to the same or different languages when nomenclature is ambiguous or imprecise. They grew out of the codes used by SIL International to identify languages in their catalogue of languages the *Ethnologue* starting with the 15th edition (Gordon 2005, but have since been adopted as a world-wide standard (http://www.ethnologue.com/codes) although SIL International continues to manage the code system (for example, changes pass through them). Because of their origin these codes are sometimes known as '*Ethnologue* codes'.

For most languages there is little difficulty in matching the language with a code. However, there are some problems, mainly having to do with how coarse or fine the underlying classification of language varieties is assumed to be. The codes are mainly aligned with the division into languages that is presented in the *Ethnologue*. In some cases this is extremely fine: Arabic is 35 distinct languages, German is 18. But in other cases it is coarse: Mandarin Chinese (cmn) is treated as a single language even though the 15th edition of the *Ethnologue* itself used to report that "Mandarin varieties of Guilin and Kunming are inherently unintelligible to speakers of Putonghua [Standard Mandarin]." When fine distinctions are made it can be hard to determine which of a number of possible codes is appropriate. When distinctions are coarse, the same code may match to two or more entries in the database. Thus, a few entries in LAPSyD have more than one code, or may share a code with another entry, as is the case of Standard Modern Greek and the Greek variety known as Grico

spoken in the Salento region of Italy, both coded ell.. The set of ISO codes is also regularly updated, so the codes for certain languages may need to be changed to keep them in conformity.

In a few instances, there is no code yet provided. These will be temporarily assigned the unused code xxx. Cuitlatec, formerly spoken in the Mexican state of Guerrero, is a current example of a language with no IS0-639 code.

Languages can be searched for by name, including alternatives, or by ISO code by clicking on the “Access languages” tab in the main menu, then “Text search”. This opens a window in which the search term can be entered, and its type (language name, code) selected. This window also allows for search by source.

2.4.*Language classification*.

For each language basic information on its language family membership is provided. In the majority of cases, the language is classified into a high-level family on the order of Indo-European, Niger-Congo, Austronesian, Sino-Tibetan or Arawakan, and then into a major sub-family, not necessarily the next lower node in a familiar tree. In families such as Indo-European and Afro-Asiatic which split at a high level into multiple recognized branches the major subfamily is the next level of the classification (e.g. Italic or Germanic; Berber or Chadic). But in families such as Austronesian and Niger-Congo where branches are heavily nested, a lower level is usually used (e.g. Oceanic in Austronesian). In many cases one or more further layers of a classification are also included, but no attempt is made to provide a full classification. There is a marked lack of consensus on the structure of most language families at both higher and lower levels, so these classification labels are only intended to serve as a rough guide to relatedness and as a possible basis for constructing samples of languages appropriately reflecting genetic diversity. Languages can be searched by family by clicking on the "Access Languages” tab in the main menu then “By Classification".

The language classification assumed is moderately conservative. Larger units that are reasonably probable are accepted even though they may be rejected by more conservative comparative linguists. These include families such as Gulf (Muskogean + Tunican), Dene-Yeniseian (Na-Dene +Yeniseian), as well as contested larger groups such as Australian and Khoisan. Proposed macro-groupings such as Austric (Austro-Asiatic + Austronesian), Niger-Saharan (Niger-Congo + Nilo-Saharan) and Ural-Altaic are not accepted, let alone hypothesized mega-groupings such as ‘Amerind’ or ‘Eurasiatic/Nostratic’. However, a user is always free to reshuffle the classification to suit their tastes. Suggested tools to compare classifications are the *Multitree*

http://new.multitree.org and *LL-Map* projects hosted by the Linguist List (http://linguistlist.org/projects/).

A special treatment is given to the languages that have sometimes been labeled Papuan. These are the languages of island South-East Asia and the South-West Pacific that do not belong to the Austronesian or Australian families. Without implying a genetic unity these languages are all labeled "Papuan" (to be read with scare quotes!) at the top level, in part because the process of sorting of these languages into genetic groups seems to be in greater flux than is the case elsewhere. A second-level affiliation, such as Trans-New Guinea, is also provided for these languages, which may correspond more closely to independent families.

Many languages are of uncertain affiliation; in LAPSyD these may be variously included in a more established grouping or left unaffiliated. For example, Japanese and Korean are shown as Altaic, a grouping accepted by some and rejected by many. Languages with no known affiliation are of two types, at least in principle. Some are quite well documented and have been the subject of concerted efforts to decide their affiliation and yet these efforts have failed to show demonstrable relationship to any other language. These are labeled Isolates, with the name of the language following. For example the Zuni language spoken in New Mexico is labeled: Isolate: Zuni. Other languages are not documented in sufficient detail (e.g. only a small vocabulary may be available) or have not been evaluated in detail from the perspective of their classification. These are labeled Unclassified. Proposed but uncertain affiliations are occasionally noted in individual language files.

2.5.*Language localization*.

The location where each language is spoken is specified in two ways. There is a brief verbal description of the area where speakers of the language are primarily resident now — or in some cases were resident before major population displacements due to colonial intrusions, etc. These descriptions are either in terms of geographical features such as river systems, islands, mountain chains or in terms of administrative and political units such as towns, departments and provinces, or both. The country is always mentioned.

In addition there is a point location, indicted by co-ordinates of latitude and longitude in decimal format, intended to mark the center of the region where the language is spoken, or its most characteristic 'habitat'. The majority of the world's languages are spoken by reasonably sedentary populations settled within a relatively small area. Assigning a single point location is not a great distortion of reality on the ground in these cases. But languages spoken by highly mobile or dispersed populations (e.g. Romani, Fulani) or by large populations spread

over a great area are of course poorly represented by a single point. In the first of these situations, the point location is based on the location where the data on the particular variety of the language described was obtained. Thus, Cherokee (chr) is placed in western North Carolina in the USA rather than in Oklahoma, as sources describing the North Carolina dialect provide the primary information. This location also better represents the historic home of the language. In the case of widespread languages a political center is chosen to represent the language's location. Thus, English is located in London, Spanish in Madrid, Russian in Moscow, Mandarin Chinese in Beijing and Hindi in New Delhi.

Assigning point locations has some advantages over attempting to map areas where languages are spoken. It avoids the difficulties of determining boundaries and dealing with overlapping ranges, and avoids giving a false impression of precision. People move around, voluntarily or under duress, and may change the language they use. The relatively abstract nature of point locations serves to remind the user of the somewhat fictional nature of all language mapping. LAPSyD's visualization tools use the point locations when plotting locations for all or selected subsets of the languages included in the database.

2.6.*Areal/genetic groups*.

Each language is assigned to one of six major areal/genetic groupings. Such groups are often used to test the generality of typological observations. For example, if patterns are repeated in each group separately, they are more plausibly universal. The areal/genetic groupings are defined first on a geographic basis. All of the languages belonging to families wholly or primarily based in a given geographic area are attributed to the area. The six groups are as follows:

1. Europe, South and West Asia.
2. East and South-East Asia.
3. Africa.
4. North America.
5. Central and South America.
6. Oceania.

The representation of the different areal groups can be seen by clicking on the "Summary" tab in the menu and then clicking on "Language areas".

Area 1 includes all the countries of Europe, including all of Russia and the Central Asian republics of the former USSR, as well as Asia Minor and Anatolia, and the Indian subcontinent. Major language families rooted in this area are Indo-European, Uralic, Altaic, Dravidian and the three Caucasian groups. A number of isolates and small families, such as Basque, Burushaski and Chukchi-Kamchatkan, also fall into this group. Creoles that are lexically

primarily Indo-European have been classified as Indo-European, and hence are included in group 1 no matter where they are spoken. Since Korean and Japanese have frequently been linked to Altaic they are also included in area 1 rather than in area 2.

Area 2 includes China, the countries of mainland South-East Asia and the islands north and west of Wallace's Line (i.e. most of Indonesia and the Philippines). The major language families in this area are Sino-Tibetan, Austro-Asiatic, Tai-Kadai, Hmong-Mien and Austronesian. Sino-Tibetan and Austro-Asiatic (Munda) languages spoken within the Indian subcontinent are included here, as are all the Austronesian languages spoken around the Pacific and the Malagasy language (plt) of Madagascar. Among smaller families in area 2 is Andamanese, since the Andaman islands though politically part of India are geographically offshore from South-East Asia. Moreover, it has recently been suggested that Andamanese might be related to Austronesian.

Area 3 is the African continent and its offshore islands and includes all languages of the four traditionally recognized major language families of Africa: Niger-Congo, Nilo-Saharan, Afro-Asiatic and Khoisan. There are increasing doubts among Africanists as to whether all the languages usually considered to belong to the first two of these are in fact related, but each contains a very large undisputed core membership. Afro-Asiatic includes all Semitic languages, some of which are spoken outside Africa, e.g. in Malta and the Middle East. Khoisan is a disputed grouping supported by some and rejected by most specialists working on these languages, but nonetheless remains a common frame of reference. There are also a small number of isolated or unclassified languages in Africa, including Hadza (hts).

Area 4 is the North American continent, which is defined as reaching to the Isthmus of Tehuantepec, so that most but not quite all of Mexico is included in the North. Many distinct language families are recognized in this area. Some of the larger families are Dene-Yeneseian (Na-Dene + Yeniseian), Algic, Iroquoian, Uto-Aztecan, Oto-Manguean and Hokan and Penutian. The latter two contain reasonably certain core membership, but many disputed extensions have been proposed. Quite numerous smaller families and isolates also exist in this area, such as Wakashan, Kiowa-Tanoan, and Kutenai. North American families generally have no members outside the strictly geographical boundaries of the area. An exception is Ket (ket), now recognized as linked to Na-Dene in the Dene-Yeniseian family. Although it is possible that the north Asian location of Ket represents the older homeland of this family, it’s members are dominantly spoken in North America, so Ket joins this area.

Area 5 includes the Americas south and east of the Isthmus of Tehuantepec as

well as the islands of the Caribbean. This division places the Yucatan peninsula in the Central and South American area and unites all languages of the Mayan family in this group, together with families such as Chibchan, Arawakan, Cariban, Tupian, Pano-Tacanan, Tucanoan and many smaller families. There are also a considerable number of South American languages which appear to be isolated or are as yet unclassifiable.

Area 6 — "Oceania" — includes the islands east and south of Wallace's Line, most importantly New Guinea, and the island-continent of Australia, but also the smaller islands of the southern and eastern Pacific and a further significant part of Indonesia (Sulawesi, and the Maluku and Lesser Sunda islands such as Flores and Timor but excluding Bali which lies west of Wallace's Line). The languages assigned to this area belong to the Australian family or to one of the various groups that have been labeled 'Papuan'. Opinions on the internal classification of the indigenous languages of Australia seem to be becoming more and more stable, but the classification of 'Papuan' languages remains very unsettled. Since the 'Papuan' language groups are often discussed together, they have all been designated as 'Papuan' followed by a suggested family affiliation. The numerous Austronesian languages spoken in this geographic region are assigned to area 2 — East and South-East Asia — since languages in the Austronesian family are primarily spoken to the west and north of Wallace's Line.

2.7.Sources consulted.

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. Occasionally an additional source that has been identified but not yet consulted is also listed together with the annotation “not seen”. 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, including recorded speech samples. In some cases a web address is provided which will link to the item cited. As described in ¶ 3 the phonological analyses presented in the sources are not necessarily accepted as given, but are modified with the intention of achieving a uniform interpretation of the facts.

3. Interpretation and standardization of descriptions

3.1.General principles

The phonological description of each language is reviewed to standardize the analyses as far as possible. The basic idea is to remove differences that have to do with choice of theoretical model or transcriptional preferences and other issues that might either create apparent rather than real differences between the

languages included, or might disguise real differences that actually exist. The goal is to represent the language in a 'concrete' fashion based on what is actually produced in a careful speech style.

The primary goal of this database is to represent the segmental and prosodic contrasts that form lexical distinctions in each language, together with basic information on phonotactics. Most languages yield fairly well to an analysis in terms of a set of contrastive segments that can be identified through the classic test of contrast in minimal pairs. These elements are generally known as the phonemes of the language. Each element identified in this way can then be characterized as possessing certain phonetic traits. These will normally be the properties that occur in the most common variant produced in a reference pronunciation. There are, of course, many decisions to be made about identifying the segments as produced in different environments with each other. In general, linguists tend to rely on phonetic similarity above all else. However, sometimes other arguments can support an identification, such as the identity of morphemes. For example, in a set of English (eng) words such as *fate*, *fatal*, *fatality* the assumption that they share the morpheme <fate> provides a basis for saying that the rather different sounds heard in utterance-final, medial post-stress and medial pre-stress positions in these words spoken in isolation are all realizations of the same element /t/. Phonetic attributes that can be reasonably derived from particular environments, such as the aspiration found in pre-stress onset position, are not considered part of the essential nature of the segment. So /t/ in English is considered a voiceless alveolar plosive (pulmonic stop) and is not marked for aspiration.

A different case can be illustrated with Bilua (blb), a 'Papuan' language of the Solomon Islands. In this language the voiced stop series is described as being prenasalized when intervocalic but plain word-initially (Obata 2003). Since prenasalization is not explicable as a reasonable consequence of simply being intervocalic, it is assumed that the voiced stops in this language are basically prenasalized. The older spelling of the language name as Mbilua suggests that in fact they probably are prenasalized in initial as well as in medial positions but prenasalization may be less salient in this position.

Following such principles, the most reasonable basic form to posit for each contrastive segment in the language is sought.

Of course, all of the classic problems of determining an inventory of segments must also be considered. These include evaluating ongoing sound changes that may be modifying the inventory, such as splits and mergers in progress and the effects of contact between language leading to the possible introduction of new sounds. Decisions must also be made about whether phonetically complex

entities should be treated as units or as a sequence of segments. All elements that are candidates to be considered as unitary affricates, diphthongs, consonants with secondary articulations, prenasalized stops, and many other types of segments traditionally accepted as potential unitary elements in the phonetic literature, are also candidates to be considered as a sequence.

In LAPSyD, unlike in UPSID, there is no marking of loan segments. For each case where a segment is known (or believed) to have entered a language because of borrowing, a judgment is made as to whether the borrowed segment is sufficiently integrated in the language. For example, English has borrowed the word *genre* from French (fra) and some speakers may pronounce this word with a nasalized vowel [ã], similar to that in the French pronunciation [ʒãʁ] (though never with a French-sounding uvular r-sound). But this pronunciation is not used by a majority, so /ã/ is not considered part of the established inventory of English. On the other hand, On the other hand French has borrowed so many words from English with the ending <–ing> that this has become a morpheme that can be added to native French roots. Hence French now has a phoneme /ŋ/. Many languages are spoken in areas where a dominant language has had a major impact — for example, virtually all indigenous languages of the Americas are subject to influence from English, Spanish (spa), Portuguese (por) or French. Descriptions will often note that certain sounds or syllabic structures only occur in, say, loanwords from Spanish. Clues are sought as to whether such items should be considered as cited from the dominant language, or as having become an integral part of the indigenous language. In order to diminish the strong homogenizing effects of such dominance, a bias against accepting introduced elements is taken as the initial stance. This, of course can be overcome when the evidence of nativization is persuasive.

All the data in LAPSyD is subject to any 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 check, for example, how many long vowels occur or to construct an idea of the syllable canon, but not all such lacunae can be filled.

3.2. Vowel inventory

The inventory of vowel nuclei recognized for the language is given in IPA transcription with the symbols laid out in a basic grid showing vowel height (vertical) and front-back dimensions (horizontal with front at the left). Rounding is shown by choice of symbols and by labeling. Other properties will be most often indicated by diacritics. Note that the symbol /a/ is used for a low

central vowel, not a front one. If the language has diphthongs, that is, dynamic nuclear vowels, these are listed beneath the vowel grid. Many linguists use the term "diphthong" for structures that consist of an approximant and a vowel (in either order). These are not necessarily syllable nuclei but may rather be CV or VC structures. The description is searched for indications as to the best analysis. For example, if a vowel + approximant rhyme precludes the occurrence of another coda consonant, as in Thai (tha), this indicates that treating the approximant as a coda itself may be the most appropriate analysis.

A commentary field titled "Vowel notes" provides space for remarks related to the vowel system and its interpretation. These may include notes on how the data reported in LAPSyD differs from the source(s) consulted, as well as on such issues as restrictions on the distribution of certain vowels, such as vowel harmony. Information missing from the source may also be noted here. The commentaries are not in any consistent format, but may sometimes contribute to understanding the way data has been interpreted.

The total number of distinct vowel nuclei is listed in the "Count Information" provided for each language. This is the sum of all vowels and diphthongs of all types. This total is often less certain than the number of basic vowel qualities.

3.3. Basic vowel count

A count of basic vowel qualities is also provided. This collapses distinctions among vowels that have the same values on the basic parameters of height, backness and rounding. Pairs of, say, oral and nasalized vowels, or long and short vowels that can be matched to each other on the basic parameters are counted just once. Diphthongs that can be considered as composed of more basic vowels do not add to this count. Distinctions of tongue root position are mapped to height differences, so they do add independently to the basic vowel count. Navajo (nav) presents a clear example of the difference between total and basic vowel counts. It has four short oral vowels, /i, e, a, o/, as well as nasalized counterparts of each of these and also long counterparts to each of the oral and nasalized short vowels. So this language has 16 total vowels, but only 4 basic vowels. The decisions on the basic vowel inventory are not always as straightforward as in the case of Navajo, but the most uniform interpretation is sought. Cross-linguistic comparisons of the vowel systems of languages are very often based on the basic rather than the total vowel inventory.

If a reliable source of information on segment frequency has been found the most frequent of the vowels is reported at the right of the display of the counts.

3.4. Consonant inventory

The consonant inventory is presented in a chart labeled with axes for manner

(vertical) and place (horizontal). Place is organized from front (at left) to back, and manner largely follows the order of degree of stricture from most closed to most open constriction. Sibilant and non-sibilant africates and fricatives appear in separate rows. Languages with clicks have a second chart for the clicks. Subsidiary distinctions, such as secondary articulations, are nested within the main axes of the grid. The conventional description of /h/ as a voiceless glottal fricative is accepted for the purposes of this database.

A count of the number of consonants recognized for the language is provided under the "Count Information". In addition, the ratio of the number of consonants to both the number of total vowels and the number of basic vowels is reported. Navajo has 34 consonants, so for this language these ratios are 2.125 (34/16) and 8.5 (34/4) respectively. As for vowels, if a reliable source of information on segment frequency has been found the most frequent of the consonants is reported at the right of the display of the counts.

A commentary field titled "Consonant notes" provides space for remarks related to the consonant system and its interpretation. As for vowels, this field may include notes on how the data reported in LAPSyD differs from the source(s) consulted, and on restrictions on distribution. This field often includes notes on the description given to coronal consonants in a source, or on ambiguities in the labeling. These commentaries are not in a consistent format, but may add some details or contribute to understanding the way data has been interpreted and where remaining uncertainties lie.

3.5.Syllable structure

The basic patterns of syllable structure are reported in several ways. At the beginning of the field titled "Syllable notes" a notation of the canonical syllable structure assumed for the language is given, using the standard notation of C for consonant and V for vowel. Items in parentheses are optional. Thus (C)V(C) means that the language allows four types of syllables with V, CV, VC and CVC structures. Common patterns allowing restricted classes of consonants in certain syllable positions may be noted with the symbols G, L and N, for Glide, Liquid and Nasal respectively. Thus C(G)V would represent a syllable canon in which an onset is obligatory and it may have the structure of just a single consonant or a consonant followed by /w/ or /j/ or a similar approximant. The notation V(V) indicates cases where there seems to be relatively free combination of vowels which nonetheless do not create separate syllabic nuclei. The notation V(:) indicates long vowels occur. The presentation of the syllable canon may be followed by a commentary noting, for example, difficulties of interpretation, the existence of other phonotactic limits on syllables, e.g. if word-medial and word-final syllables differ in structure, and the field may provide examples of the different syllable patterns. Incompleteness of

information may also be noted here.

In order to make canonical syllable structure searchable two fields are provided, one (‘Canonical Form’) containing the most fully elaborated syllable canon in simplified form using only C and V as well as : in case of long vowels, and the other (‘Syllabic Restriction’) noting where there are substantial restrictions on the consonants that occur in a given position (noted by *). Many languages have smaller sets of permitted coda consonants than onset ones but a coda restriction is only shown in this field when the number is decidedly limited.

Each language is also assigned to one of three categories for its syllabic complexity ("Syllcat") corresponding to those in the WALS database (http://wals.info/feature/12A). Languages that allow nothing more elaborate than a CV syllable as classed as having Simple syllable structure; those which allow either a common type of two-consonant onset, such as CG or CL, or allow a single consonant in coda position, or allow both of these are classed as Moderately Complex. Languages that have less common onset clusters, such as two obstruents or three or more consonants, or which have any clusters in the coda are classed as having Complex syllable structure.

Numeric values are also given for the maximal degree of elaboration of the Onset, Nucleus and Coda elements separately, and these are summed to give an overall Syllable Index. Onset values are 0 for maximal one-consonant onset (since the CV syllable type is taken to be universal), 1 for common CC onsets (such as CG, CL), 2 for less common 2-consonant onsets, and 3 for maximally 3 or more consonants in onset. Nucleus scores are 1 for single-mora nuclei as the maximum, and 2 for bimoraic (or potentially longer) nuclei. Coda scores are 1 for maximally a single consonant, 2 for two consonants and 3 for 3 or more consonants permitted in coda. The summed Syllable Index thus ranges from 1 for a languages such as Yoruba (yor) with maximal CV syllables to 8 for a language such as English (eng) which permits elaborate onset and coda clusters. In addition data is imported from the relevant WALS chapters by Goedemans & van der Hulst on whether the language has been interpreted by these authors as having Fixed stress location (http://wals.info/chapter/14) or stress placement that is affected by syllable weight parameters (http://wals.info/chapter/16). These interpretations may differ from that preferred in LAPSyD.

3.6. Tone system

The commentary field "Tone" briefly summarizes what is known about any system of lexical or grammatical tone contrasts that the language has. For languages known or presumed to have no tone system this fact is also noted.

Common notations such as H, M and L for High, Mid and Low are often used in this field, but the numeric notation used by many linguists working on Asian languages (5 = High, 1 = Low) is also sometimes employed when describing these languages.

In the "Tonecat" field the complexity of the tone system is expressed by use of one of four category labels: None, Simple, Moderately Complex, Complex. Languages with a basic two-way contrast (which may include a limited use of downstep) are classed as having a Simple tone system, those with three contrasts are Moderately Complex. Complex tone systems have four or more contrasting tones. A few languages are noted as having a Marginal tone system. These cases are of two types: languages where tone distinctions are said to be relevant to only a small part of the lexicon, and those where the tone distinctions might be subsumed under an accentual contrast.

3.7.Stress (accent) system

The commentary field "Stress" contains summary information on the presence and role of any reported stress distinction in the language. The principal intent is to report whether stress plays a role in distinguishing lexical (or grammatical) forms. The "Stresscat" field categorizes languages into three groups according to the role of stress: None, Minimal, Lexical. Those languages reported or presumed to make no perceptible differences in stress level between syllables are labeled "None". Languages in which there are noticeable differences in stress level, but the placement of stress in lexical items is predictable (or very largely so) are classed as having a Minimal role for stress. Languages in which stress placement distinguishes (at least some) lexical forms or is otherwise unpredictable are labeled as having a Lexical role for stress.

It is not always straightforward to decide when to characterize a language as having a system of tonal contrasts or one of accentual contrasts as the two are not sharply distinguished in reality, and some languages clearly have both. In LAPSyD a forced choice is made as to whether any given language has tone, stress or both but the notes help to identify where different opinions might be justifiable.

4. **The feature description of segments**

All of the vowel and consonant segments referenced in the database are given a unique 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, patterns of complementary distribution and other properties of the inventories at the featural level.

4.1. *General principles*

The features used to define the segments that are catalogued in LAPSyD are based on traditional phonetic terminology, such as that embodied in the charts of the International Phonetic Alphabet including all the often basic distinctions that are made by diacritical marks. More elaborate and precise classifications, such as that presented in *Sounds of the World's Languages* (Ladefoged and Maddieson 1996), cannot be systematically used due to the lack of precision in a good number of the available descriptions. For example, the standard term 'retroflex' is used for at least three distinguishable articulatory postures, which could be described as sublaminal post-alveolar, apical post-alveolar, and laminal post-alveolar. In many publications where the label 'retroflex' is used, there is no indication as to which of these might be the articulation used in the language.

However, 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 segment in a language's inventory of consonants and vowels has a distinct feature representation. This representation is the same as that assigned to segments in other languages which are judged to share the same classificatory characteristics. In three cases, explained more fully in the feature enumeration below, there is such a frequent lack of clarity in source descriptions that special features encoding the ambiguity are used. These features are *unspecified coronal place*, *unspecified rhotic* and *unspecified mid* (vowel height). Segments are assigned one of these features if it is unclear which particular coronal place, rhotic type or vowel height they typically present. Segments bearing one of these features do not form a coherent class; rather, it is uncertain which class they should be allocated to. These ambiguities most often arise when just a list of symbols is given without specific definitions of their intended value. Sometimes an informed guess can be made based on knowledge of traditions of scholarship for a particular language area or family, but in many cases ambiguity remains. There are also instances where segments are described using inherently ambiguous phonetic terms, such as *denti-alveolar* or *vibrant*.

Establishing a consonant or vowel inventory presupposes an agreed segmentation. Decisions must often be made as to whether a particular consonantal pattern represents a single complex segment or a sequence of two or more segments. Similar questions arise with vowels where alternative analyses might posit unitary diphthongs, or sequences of two independent vowels, or combinations of a vocalic approximant and a vowel. In such cases it is often distributional patterns that provide the best support for the choice of analysis. For example, [ts] and [tʃ] have very different distributions in English. [tʃ] occurs in syllable onsets where other stop+fricative sequences do not occur.

[ts] occurs in syllable codas, where other combinations of stop+fricative do occur, especially other non-homorganic voiceless stop + [s] combinations (as in *lax* and *lapse*), and moreover the [s] in this position often represents a separate morpheme (as in *lacks* and *laps*). [tʃ] thus has a unitary character, whereas [ts] is naturally interpreted as a sequence of two separate consonants. As for vowels, a language that seems to allow all (or almost all) possible combinations of its simple vowels, such as Lavukaleve (lvk), is more readily judged to have sequences of independent vowels than to have a large inventory of diphthongs. Since in this language either the first or the second vowel in a VV sequence might bear stress, the preferred analysis that each vowel forms a syllable. The fields in the database for comments on the vowels and consonants provide for brief discussion of such issues in individual languages. It is not uncommon for a different interpretation to be preferred in LAPSyD to the one offered in the literature.

In the following sections (¶ 4.2-4.6) the full feature set available to characterize the segments in LAPSyD is presented

4.2.*Consonant Features* **(some also applicable to vowels)**

4.2.1. **Features that create separate rows**

Airstreams

pulmonic (= pulmonic egressive)
ejective (= glottalic egressive)
implosive (= glottalic ingressive)
click (= velaric ingressive)

The default airstream for speech is provided by the lungs. Thus almost all segments have the feature *pulmonic*. Segments assigned this feature have *only* a pulmonic airstream. Pulmonic airflow can be briefly interrupted or modified by constriction and vertical displacement of the larynx as in implosive stops and ejective stops, affricates and fricatives, or by expansion of the oral cavity, as in clicks. Segments whose prototypical production is as an ejective or implosive may be produced without sufficient larynx movement to actually create outward or inward airflow, but they are still classified as *ejective* or *implosive*. Stop segments described in the sources as 'glottalized' and notated with an apostrophe (e.g. /k'/ or /k̓/) are usually interpreted as ejectives, and segments described as 'pre-glottalized' and notated with a voiced stop symbol and some mark of glottalization (e.g. /ʔb/, /ˀb/ or /'b/) are usually interpreted as implosives. In both cases, this is especially likely if related or neighboring languages are known to have ejectives or implosives. Implosives are most often produced with voicing — at least at their release — due to pulmonic airflow through the descending glottis, but in some languages it has been argued that there are implosives of two types,

with and without voicing. So-called ‘voiceless’ implosives maintain a complete glottal closure. For this reason a voicing feature is always assigned to implosive segments. In this context *voiceless* must be understood as implying a fully closed glottis, rather than open vocal folds. In clicks the back closure release always involves the pulmonic airstream, and there may be voicing and/or nasal airflow due to pulmonic air during the hold of the front oral closure. However, clicks are not assigned the feature *pulmonic* even in these cases. Instead, these properties are indicated by voicing features and the feature *nasalized*, as appropriate. Clicks can also be produced with an ejective release of the back closure. Thus *ejective* and *click* can co-occur on a single segment. This is the only possible combination of two features from this set. There is a further discussion of some of the special considerations that apply to clicks in section 4.5 below.

Manners

stop
affricate
fricative
nasal
trill
tap/flap
unspecified rhotic
approximant

All features in this set are mutually exclusive. The manner features primarily represent the degree of stricture required for a segment's production: *stop* and *nasal* for full closure, *fricative* for narrow approximation, *approximant* for open approximation. *Affricates* combine a stop phase with a fricative release. Note that the feature *nasal* is only assigned to purely nasal consonants such as /m, n, ŋ/ produced with complete oral closure. The features *trill* and *tap/flap* describe consonants with intermittent or very brief closures. Trilling is aerodynamically-driven and occurs only within critically narrow limits, so trills regularly vary with productions in which something other than an actual trill occurs. If a segment is reported to be produced with multiple contacts in some instances it is usually assumed that the reference pronunciation is a trill. Taps and flaps can be distinguished as motions of the moving organ orthogonal to or parallel to the contacted surface respectively, but this distinction is not reliably adhered to in most descriptions. Hence these classes of segments are collapsed in LAPSyD. The remaining feature in this set, *unspecified rhotic*, is assigned to segments which are “some kind of r-sound” but for which the sources available on the language do not permit a more complete description to be made. Often no more that the

symbol <r> may be provided, or an ambiguous term such as 'vibrant' may be used. These segments would be assigned to one of the *trill*, *tap/flap* or *approximant* categories if the correct assignment was known. They do not constitute a distinct category on their own. Rather, the *unspecified rhotic* feature is a device to avoid misrepresentation in the face of ignorance.

Secondary Sources

sibilant
whistled

The two secondary source features describe factors that shape the acoustic spectrum of fricatives (and affricates) downstream from the primary constriction. In *sibilants* air is channeled through a primary constriction so that it strikes a downstream obstacle (typically the back of the upper teeth) creating strong high-frequency turbulence. The feature *whistled* is added for fricatives and affricates that have a configuration of the lips that adds a whistle-like resonance to the sound, as well as filtering higher frequencies.

Escape/Release/Approach features

lateral
nasalized (also for vowels*)*
prenasalized
pre-stopped
trilled-release

This set of features specify non-default escape paths for the airflow, the default being oral only and central, as well as modifications taking place in the supra-laryngeal vocal tract of the approach or release of obstruents. *Nasalized* vowels or consonants, except for nasalized clicks, have simultaneous airflow through the oral and nasal cavities. In nasalized clicks air flows out through the nasal passage while the click mechanism blocks airflow through the oral cavity.

4.2.2. Features that create separate columns:

Places

bilabial
labio-dental
linguo-labial
dental
alveolar
unspecified coronal place
palato-alveolar (= laminal post-alveolar)
retroflex (= apical post-alveolar)
palatal

velar
uvular
pharyngeal
epiglottal
glottal

(known combinations: placed in a new column following the column of the non-labial place concerned)
bilabial+velar
bilabial+alveolar
bilabial+palato-alveolar (or retroflex) (as in Yélî Dnye)
bilabial+palatal

Secondary articulations

labialized
palatalized
velarized
pharyngealized (also for vowels)
epiglottalized (for 'sphincteric' vowels)

4.3.*Voicing Properties*

On consonant charts these create separate rows

For vowels and consonants

voiceless
voiced
breathy voiced
laryngealized

For consonants *(in combination with 'voiceless' or 'voiced')*

aspirated
pre-aspirated
breathy-release
pre-voiced (transcribed with a symbol sequence. e.g. /dt/)

4.4.Duration Properties

On consonant charts these create separate rows

For vowels and consonants *('normal' length is unmarked)*

long
overshort

4.5.Some special considerations for clicks

Clicks are among the most complex of human speech sounds and many issues remain to be clarified in their description. Clicks of a given language are often given very dissimilar descriptions by different authors. This section explains some of the decisions reached in trying to harmonize their characterization.

All clicks require closures to be made at two locations within the oral cavity; each of these closures may be released abruptly or gradually, in the latter case creating an affricated release of the front or back closure. The click mechanism itself takes place entirely within the oral cavity so the larynx is free to adopt any configuration from complete closure (glottal stop) to a fully open voiceless position, and can also operate as an airstream iniator creating an ejective release of the back closure. The velum may be raised or lowered to prevent or permit nasal airflow. The timing of actions of the larynx and velum can vary both with respect to the oral articulations of the click and with respect to each other.

Place features: The front click closure locations are specified with the same place features as other consonants. For now, the view that back closures can be contrastively *velar* or *uvular* is accepted, although the difference has perhaps more to do with the audibility of the back release than its actual location. (When the back release is delayed its place of articulation is readily perceived as further back than where the back closure is formed at the onset of a click. The closure is retracted during the closure phase.) Hence all clicks are assigned two places of articulation. Clicks with a simple glottal closure released after the click are also given the place feature *glottal*. This is different from an ejective release.

Release features: clicks are assigned the feature *affricate* if the **back** closure is released with frication. Clicks lacking this feature have a stop-like back release which may be followed by aspiration. Sustained voiceless nasal air-flow after a click release is described by the combination of features *nasalized* and *aspirated*. In most languages with clicks if the click is bilabial, dental or alveolar lateral the front closure is released slowly creating some frication. Front closures of post-alveolar and palatal clicks are generally released abruptly. However, some languages may have abrupt lateral click releases or slow palatal click releases. This property was indicated in UPSID (Maddieson 1984) with a feature *affricated click* for the slowly released cases. Since no language seems to make a distinction at the same place of articulation between these two kinds of front release, this feature is not retained in LAPSyD. Some languages, such as |Nuu (ngh) distinguish ejective click releases with and without affrication of the release.

4.6. *Vowel Features*

Major classes

vowel
diphthong
triphthong

Vowel height

suprahigh
high
higher mid
unspecified mid
lower mid
low

Vowel backness

front
central
back

Vowel lip position

unrounded
rounded
lip-compressed

Subdivision of vowel height space

raised
lowered

Subdivision of vowel backness space

fronted
retracted

Ordering features for diphthongs/triphthongs

rising
lowering
backing
fronting
rounding
unrounding
nasalizing
devoicing

4.7.Superordinate class features

The following superordinate features are also used. These are designed to facilitate searches for familiar classes of segments otherwise only specifiable by a complicated set of feature specifications:

obstruent (all stops, affricates and fricatives)
liquid (rhotics and voiced lateral approximants)

5. **Transcription**

The transcription provided for all consonants and vowels in LAPSyD follows the usage set out in the *Handbook of the IPA* in most regards, within the limits allowed by the desire to standardize across the sources used. The following deviations are the principal differences from conventional IPA transcription.

As noted in the discussion of Vowel Inventory (¶ 3.2), /a/ is systematically used for a low central unrounded vowel, not a front one. To indicate a low front vowel the fronting diacritic is added, i.e. /a̟/. Mid vowels that are not specified as being in the higher mid or lower mid range are represented with the higher mid symbols and single quotation marks, e.g. /'e', 'o'/.

Consonants articulated in the coronal region are often rather imprecisely described in the sources. Those specifically noted as dental are transcribed with the dental diacritic (e.g. /t̪, n̪, s̪/, etc), except that /θ, ð/ are taken to explicity represent dental articulations and require a retraction diacritic to indicate further back articulation. Symbols such as /t, n, z/ without any diacritic are used to represent alveolar articulations. In those cases where a source uses a symbol for a coronal articulation without further specifying place of articulation, the transcription in LAPSyD uses single quotation marks to indicate "unspecified coronal place" (abbreviated "un-cor" on consonant charts), e.g. /'t', 'l', 'n'/. Rhotic segments whose manner of articulation is not defined (e.g. as a trill or an approximant) are transcribed with doubled letter r (i.e. as /rr/). A rhotic unspecified for manner might also be unspecified for place, in which case the transcription is /'rr'/.

6. **References**

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

Dryer, Matthew S. & Haspelmath, Martin (eds.). 2013. The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig. Online at http://www.wals.info/.

International Phonetic Association. 1991. *Handbook of the IPA*. Cambridge University Press, Cambridge.

Gordon, Raymond G., Jr. (ed.), 2005. Ethnologue: Languages of the World, Fifteenth edition.

SIL International, Dallas. Current online version: http://www.ethnologue.com/.
Maddieson, Ian. 1984. *Patterns of Sounds*. Cambridge University Press, Cambridge. Paperback edition, 2009.