sciendo Yearbook of the Poznań Linguistic Meeting 4 (2018), pp. 21–53
DOI: 10.2478/yplm-2018-0002

A cross-linguistic database of phonetic transcription systems

Cormac Anderson[1], Tiago Tresoldi[1], Thiago Chacon[2], Anne-Maria Fehn[134], Mary Walworth[1], Robert Forkel[1] and Johann-Mattis List[1]*
[1]*Max Planck Institute for the Science of Human History, Jena;* [2]*Universidade de Brasilia;* [3]*Goethe University, Frankfurt;* [4]*CIBIO/InBIO: Research Center in Biodiversity and Genetic Resources, Vairão, Portugal*
* *list@shh.mpg.de*

Abstract
Contrary to what non-practitioners might expect, the systems of phonetic notation used by linguists are highly idiosyncratic. Not only do various linguistic subfields disagree on the specific symbols they use to denote the speech sounds of languages, but also in large databases of sound inventories considerable variation can be found. Inspired by recent efforts to link cross-linguistic data with help of reference catalogues (Glottolog, Concepticon) across different resources, we present initial efforts to link different phonetic notation systems to a catalogue of speech sounds. This is achieved with the help of a database accompanied by a software framework that uses a limited but easily extendable set of non-binary feature values to allow for quick and convenient registration of different transcription systems, while at the same time linking to additional datasets with restricted inventories. Linking different transcription systems enables us to conveniently translate between different phonetic transcription systems, while linking sounds to databases allows users quick access to various kinds of metadata, including feature values, statistics on phoneme inventories, and information on prosody and sound classes. In order to prove the feasibility of this enterprise, we supplement an initial version of our cross-linguistic database of phonetic transcription systems (CLTS), which currently registers five transcription systems and links to fifteen datasets, as well as a web application, which permits users to conveniently test the power of the automatic translation across transcription systems.
Keywords: phonetic transcription; phoneme inventory databases; cross-linguistically linked data; reference catalog; dataset.

1. Introduction

Phonetic transcription has a long tradition in historical linguistics. Efforts to design a unified transcription system capable of representing and distinguishing all

the sounds of the languages of the world go back to the late 19th century. Early endeavours included Bell's Visible Speech (1867) and the Romic transcription system of Henry Sweet (1877). In 1886, Paul Passy (1859–1940) founded the Fonètik Tîtcerz' Asóciécon (Phonetic Teachers' Association), which later became the *International Phonetic Association* (see Kalusky 2017: 7f). In contrast to writing systems targeted at encoding the speech of a single language variety in a visual medium, phonetic transcription aims at representing different kinds of speech in a unified system, which ideally would enable those trained in the system to reproduce foreign speech directly.

Apart from the primary role which phonetic transcription plays in teaching foreign languages, it is also indispensable for the purposes of language comparison, both typological and historical. In this sense, the symbols that scholars use to transcribe speech sounds, that is, the *graphemes*, which we understand as sequences of one or more glyphs, serve as *comparative concepts*, in the sense of Haspelmath (2010). While the usefulness of phonetic transcription may be evident to typologists interested in the diversity of speech sounds (although see critiques of this approach to phonological typology, i.a. Simpson 1999), the role of unified transcription systems like the *International Phonetic Alphabet* (IPA) is often regarded as less important in historical linguistics, where scholars often follow the algebraic tradition of Saussure (1916, already implicit in Saussure 1878). This emphasises the *systematic aspect* of historical language comparison, in which the distinctiveness of sound units within a system is more important than how they compare in substance across a sample of genetically related languages. If we leave the language-specific level of historical language comparison, however, and investigate general patterns of sound change in the languages of the world, it is obvious that this can only be done with help of comparable transcription systems serving as comparative concepts.

Here, we believe that use can be made of cross-linguistic reference catalogues, such as Glottolog (http://glottolog.org, Hammarström et al. 2017), a reference catalogue for language varieties, and Concepticon (http://concepticon.clld.org, List et al. 2016), a reference catalogue for lexical glosses taken from various questionnaires. Both projects serve as standards by linking metadata to the objects they define. In the case of Glottolog, geo-coordinates and reference grammars are linked to language varieties (*languoids* in the terminology of Glottolog), in the case of Concepticon, lexical glosses taken from questionnaires are linked to *concept sets*, and both *languoids* and *concept sets* are represented by unique identifiers to which scholars can link when creating new cross-linguistic resources. We think that it is time that linguists strive to provide simi-

lar resources for speech sounds, in order to increase the comparability of phonetic transcription data in historical linguistics and language typology.

2. Phonetic transcription and transcription data

When dealing with phonetic transcriptions, it is useful to distinguish *transcription systems* from *transcription data*. The former describe a set of symbols and rules for symbol combinations which can be used to represent speech in the medium of writing, while the latter result from the application of a given transcription system and aim to display linguistic diversity in terms of sound inventories or lexical datasets. While transcription systems are *generative* in that they can be used to encode sounds by combining the basic material, transcription data are *static* and fixed in size (at least for a given version published at a certain point in time). Transcription data have become increasingly important, with recent efforts to provide cross-linguistic accounts of sound inventories (Moran et al. 2014; Maddieson et al. 2013), but we can say that every dictionary or word list that aims at representing the pronunciation of a language can be considered as transcription data in a broad sense.

In the following, we give a brief overview of various transcription traditions that have commonly been used to document the languages of the world, and then introduce some notable representatives of cross-linguistic transcription data. Based on this review, we then illustrate how we try to reference the different practices to render phonetic transcriptions comparable across transcription systems and transcription datasets.

2.1. Phonetic transcription systems

When talking about transcription systems, we are less concerned with actual orthographies, which are designed to establish a writing tradition for a given language, but more with scientific descriptions of languages as we find them in *grammars*, *word lists*, and *dictionaries* and which are created for the purpose of language documentation. Despite the long-standing efforts of the International Phonetic Association to establish a standard reference for phonetic transcription, only a small proportion of current linguistic research actually follows IPA guidelines consistently.

2.1.1. The International Phonetic Alphabet

The *International Phonetic Alphabet* (IPA 1999, IPA 2015), devised by the *International Phonetic Association*, is the most common system of phonetic notation. As an alphabetic system, it is primarily based on the Latin alphabet, following conventions that were oriented towards 20th century mechanical typesetting practices; it consists of *letters* (indicating "basic" sounds), *diacritics* (adding details to basic sounds), and *suprasegmental markers* (representing features such as *stress*, *duration*, or *tone*). The IPA's goal is to serve as a system capable of transcribing all languages and speech realisations, eventually extended with additional systems related to speech in a broader sense, such as singing, acting, or speech pathologies. The IPA has been revised multiples times, with the last major update in 1993 and the last minor changes published in 2005.

2.1.2. Transcription systems in the Americas

In the Americas, although IPA has become more prevalent of late, there is only a minimum level of standardisation in the writing systems used for the transcription of local languages. While in North America most of the transcription systems of the twentieth century generally comprised different versions of what is generally known as the *North American Phonetic Alphabet* (NAPA, Pullum and Laduslaw 1996[1986]), in South America the picture is murkier. Although Americanist linguists have occasionally tried to harmonise the transcription systems in use (Herzog et al. 1934), we find a plethora of local traditions that have been greatly influenced by varying objectives, ranging from the goal of developing practical orthographies (often with an intended closeness to official national language orthographies), via the desire to represent phonemic generalisations in transcriptions, up to practical concerns of text production with typewriting machines (Smalley1964).[1] As a result, it is extremely difficult to identify a common Americanist tradition of phonetic transcription.

[1] Other kinds of adaptations involved modification of standard symbols such as the use of "stroke" in some letters representing stops in order to create a grapheme for a fricative sound lacking in the Latin based typography (e.g., ‹ꝑ› for voiceless bilabial fricative [ɸ], ‹ꝺ› for dental voiced fricative).

2.1.3. Transcription systems in African linguistics

Attempts to standardise the transcription of previously unwritten African languages with Latin-based writing systems date back to the middle of the 19th century (Lepsius 1854). In 1928, a group of linguists led by Diedrich Westermann (1875–1956) developed what came later to be known as the *African Alphabet*, an early attempt to enable both practical writing and scientific documentation of African languages with a minimal number of diacritic characters (International Institute 1930). In subsequent years, the system gained popularity among linguists and eventually served as the basis for the *African Reference Alphabet* (ARA, UNESCO 1978; Mann and Dalby 1987). Despite their relative success, most transcription systems and practical orthographies in use today are *mixed systems*, which inherit different parts from the IPA and the ARA, as well as alphabets of former colonial languages, alongside idiosyncratic elements. Although some areas developed regional conventions, languages with similar phoneme inventories may still be transcribed with widely diverging systems.[2]

2.1.4 Transcription systems in the Pacific

Among Oceanic languages, transcription conventions are extremely varied and are frequently based on regional orthographic conventions or the preferences of the respective linguists. In West Oceania, there is an increasing use of IPA in recent linguistic descriptions, however most existing descriptions are highly inconsistent, particularly when it comes to features that are typologically rare.[3] While Polynesian languages arguably maintain more straightforward phonological systems than their westerly cousins, they have been described with equal ambiguity. The various transcriptions include *outdated conventions*, *regional orthographic conventions*, and *individual linguists' inventions*. These have result-

---

[2] For instance, while most "Khoisan" (cf. Güldemann 2014) and Bantu languages of Southern Africa follow the African Reference Alphabet in transcribing clicks with Latin letters, linguistic treatments tend to use the IPA (following suggestions by Köhler et al. 1988). For example, the palatal click is indicated by ‹tc› in the first case and by ‹ǂ› in the second.

[3] For example, the *linguo-labial stop* of some Vanuatu languages has been described using an apostrophe following the labial ‹p'› (Lynch 2016), by using a subscript seagull diacritic under the labial ‹p̼› (Dodd 2014), and by using a subscript turned-bridge diacritic under the labial ‹p̪› (Crowley 2006a); the *doubly articulated labio-velar stop* in Vurës (Banks Islands) has been described as ‹p͡ʷ› (Malau 2016), whereas in the Avava language of Malekula, it has been transcribed with a tilde over the labial ‹p̃› (Crowley 2006b).

ed in highly ambiguous representations that easily lead to incorrect interpretations of the data, especially when being used by comparative linguists who are not familiar with the traditions.[4]

2.1.5 Transcription systems in South-East Asian languages

South-East Asian languages have a number of features that lend themselves to idiosyncratic phonetic transcription. A prominent example is tone, for which most scholars tend to prefer superscript or subscript numbers (e.g., ‹^{35}›) instead of the iconic IPA tone letters (‹˧˥›) originally designed by Chao (1930). Since scholars also use superscript numbers to indicate phonological tone (ignoring actual tone values) tone assignment can be easily confused. In addition to the transcription of tone, many language varieties have some peculiar sounds, which are not easy to be rendered in IPA and are therefore often transcribed with specific symbols common only in SEA linguistics.[5] Although especially younger field workers tend to transcribe their data consistently in IPA, we find many datasets and textbooks employing older versions of the IPA.[6]

2.1.6 Summary of transcription systems

Designing and applying phonetic transcription systems is not an easy enterprise, especially in cases where the goal is to provide a global standard. When com-

---

[4] Examples include, among others: (1) characters associated with a given sound being used to represent an entirely different sound (‹h› used for the *glottal stop*, Tregear 1899; ‹y› used for [ð], Salisbury 2002); (2) one character being used to represent various sound qualities (‹g› used for the *velar nasal* in Tregear 1899, and the *voiced uvular stop* in Charpentier and François 2015); (3) diacritics on vowels ambiguously used to indicate duration (Stimson and Marshall 1964) or *glottal stops* (Kieviet 2017).

[5] Among these are the symbols ‹ɿ› and ‹ʅ›, which are commonly used to denote vowels pronounced with friction. They could be transcribed as syllabic sibilant fricatives [z̩] and [ʐ̩], respectively, but given the problems of readability with these symbols, as well as the relative frequency of these sounds across Chinese dialects and in other Sino-Tibetan languages, scholars continue to use the symbols ‹ɿ› and ‹ʅ›.

[6] The most prominent difference is the usage of ‹ʻ› as an aspiration marker [ʰ], which can be found in many sources (Beijing Daxue 1964), reflecting an older IPA standard which is also still in use in Americanist transcription systems and occasionally still taught in recent textbooks on Chinese linguistics (see, for example, Huáng and Liào 2002). Contrast this with the frequent use of the same symbol to represent ejectives in other traditions.

paring the particular problems of transcription systems and transcription practice in different parts of the world, one can identify many similar obstacles that linguists face when trying to preserve speech in writing. The most prominent ones include (a) the influence of the orthography of the dominant language (in many parts of the world the colonial language of the oppressors), (b) a tendency to favour tradition over innovation (which results in many practices that were once considered standard now having been abandoned), (c) specific challenges in transcribing local language varieties with the material provided by the standard, (d) systemic (phonological) considerations which would entice linguists to favor symbols which reflect the phonology of the language varieties under question more properly, and (e) technical considerations (as transcription systems devised up until the mid-20th century were forced to consider the limitations of mechanical typesetting).[7] While these technical considerations should have now become largely obsolete with the introduction of the Unicode standard, this is not always the case. Judging from practical experience it is obvious that Unicode has made many things a lot easier, but since the majority of linguists are less acquainted with questions of computation and coding, the problem of typesetting is still an important factor in linguistic transcription practice.

2.2. Transcription data

In addition to transcription systems as they are used by scholars and teachers, a number of datasets offer transcription data. Usually these datasets represent typological surveys of phoneme inventories (Maddieson et al. 1984; Maddieson et al. 2013; Moran et al. 2014; Ruhlen 2008). Originally they are taken from grammatical descriptions of the languages of the world and also tend to contain an introduction into the typical sound systems of the languages under investigation. Another type of frequently available transcription data (in the sense of fixed sets of sounds which are provided in the form of transcriptions) are feature descriptions of individual collections of speech sounds which can range from single-language descriptions (Chomsky and Halle 1968), up to large collections directed towards cross-linguistic, computer-assisted applications (Mortensen 2017).

[7] This includes the IPA itself, which has many glyphs that are rotated versions of letters, e.g. IPA (1912). Further, restrictions in the early days of computing led to limited by encoding schemes such as ASCII (which led to the development of ASCII representations of IPA, such as X-SAMPA).

In a broader sense, all data collections that provide *metadata* for a given set of sounds can be qualified as transcription data. When applying this extended definition of transcription data, we can think of many further examples, including diachronic datasets of sound change (Kümmel 2008, Index Diachronica), interactive illustrations of speech sounds (Multimedia IPA chart, Wikipedia), or lexical datasets that offer phonetic transcriptions (List and Prokić 2014).

2.3. Comparability of transcription systems and data

When dealing with transcription systems and transcription data, linguists face several problems. Some of these are problems of a practical nature, which we explore further below, while others are of a theoretical nature, and touch upon long-standing issues in phonology and phonetics, and the relationship between the two. Among these theoretical problems, are those of *commensurability*, of *context*, and of *resolution*.

In spite of frequent attempts to compare phonemic inventories in phonological typology (Dryer and Haspelmath 2011; Maddieson 1984) these efforts are beset by serious difficulties. The classical structuralist treatment of the phoneme considers it to be a *relational entity* (Trubetzkoy 1939), the value of which is dependent on its place with respect to other phonemes within a system. In this understanding, the phonemes of one language are not commensurate to those of another language: it is only as a member of a system that a phoneme finds its value. This critique is taken up by Simpson (1999) who argues that the allophone replaces the phoneme in large databases, thereby reducing "the phonemic system of a language to a small, arbitrary selection of its phonetics". Although this problem cannot really be resolved, we note that different phonological databases have attempted to address it in different ways. In LAPSyD (Maddieson et al. 2013), the symbols chosen for the phonemes are often frequently occurring ones, abstracting away from too much phonetic detail. In PHOIBLE (Moran et al. 2014), on the other hand, phonemes are often transcribed with great phonetic detail, with numerous diacritics. While at first glance the latter approach might appear preferable, as it gives more information, it runs into serious difficulties, given Simpson's critique above.

The crux of this problem is that the realisation of a given phoneme depends considerably on *context*. For example, the German stops typically transcribed /b/, /d/, and /g/ are pronounced *voiceless* when in final position, whereas between vowels they are pronounced with voice. In European Spanish, while the

voiced stops /b/, /d/ and /g/ occur with the phonetic values [b], [d], and [g] in initial position, elsewhere they are more often pronounced as fricatives [β], [ð], and [ɣ]. It is not clear, in such cases, which set of symbols should be used, and even if a principled decision could be made (e.g. based on frequency, Bybee 2001), a great loss of information is involved in choosing one symbol over the other – it is equally misleading to characterise Spanish as a language without voiced stops or as a language without voiced fricatives. Such difficulties are not only of relevance in phonological typology, but can have serious repercussions in historical linguistics as well. To take an example, linguists typically transcribe two series of stops in Scottish Gaelic – aspirated /pʰ/, /tʰ/, and /kʰ/ and unaspirated /p/, /t/, and /k/. In Modern Irish, on the other hand, the convention is to transcribe rather voiceless /p/, /t/, and /k/ and voiced /b/, /d/, and /g/. In reality, however, the voiceless stops of Irish are also aspirated, and the voiced ones are only passively voiced, i.e. it is an "aspirating" language in the parlance of laryngeal typology (Honeybone 2005). The difference between these two very closely related languages lies solely in the fact that in Irish there is perhaps a greater tendency to passively voice the second series. To a naïve historical linguist, however (or indeed, to an even more naïve algorithm), this minor difference would seem a highly significant one, and would require the postulation of entirely spurious sound changes ("deaspiration" and "voicing" of the two Irish series, for example) to account for the difference.

This last example leads to a further difficulty: the level of *resolution* of the different transcription datasets available varies widely. Sapir (1930) gives an extremely detailed account of the phonological system of Southern Paiute, very rich in phonetic detail. However, in our only description of the closely related language Chemehuevi (Press 1980) there is a comparative paucity of discussion of phonetic particulars. This is not to criticise her grammar (indeed one could make exactly the opposite statement about the quality of the syntactic description in her grammar and Sapir's),[8] but rather to recognise that these two sets of transcription data have a very different level of resolution. Obviously, there are great difficulties inherent in comparing datasets of differing levels of resolution: *absence of evidence* (e.g. in some phonetic particular of Chemehuevi) does not equate to *evidence of absence*. Our degree of knowledge about the phonetics

[8] One might suggest that one of the reasons for which Press did not go into great detail on the phonetics of this language was because Sapir had already provided an extremely in-depth account of a very closely-related idiom, and thus comparatively less was known about the syntax than the phonetics of this language cluster.

and phonology of the languages of the world varies greatly, from practically nothing to voluminous descriptions detailing small sociolectal, dialectal, and idiolectal divergences.

One might ask then, given these difficulties we recognise, what the value of this enterprise is. We believe that notwithstanding these theoretical difficulties, some practical progress can still be made. Given that transcription systems are rarely standardised in a rigid manner, and allow for a certain amount of freedom of choice, scholars have come up with many ad-hoc solutions, which are reflected in specific traditions that have developed in different sub-fields of comparative linguistics. As we have seen in Section 2.1, in different linguistic traditions there are various particularities in the representation of sounds in a written medium. Scholars are usually aware of these differences in their field of expertise, but when it comes to global accounts of phonetic and phonological diversity, the particularities of the different traditions may easily introduce errors into our analyses. A great number of the practical difficulties encountered in comparative studies arise not from the broader theoretical problems outlined above, but from exactly these idiosyncrasies of tradition or personal taste. In some cases, different linguists represent sounds that are fundamentally the same in different ways (see, for instance, the examples from Pacific languages in Section 2.1.4). Convenience also plays a role here: as it is inconvenient to write a superscript ‹^{h}› to mark aspiration of a stop, scholars often just use the normal ‹h› instead, assuming that their colleagues will understand, when reading the introduction to their field work notes or grammars.[9] An ‹h› following a stop, however, does not necessarily point to aspiration in all linguistic traditions. In Australian linguistics, for example, it often denotes a laminal stop (Dench 2002).

Further problems that scholars who work in a qualitative framework may not even realise arise from the nature of Unicode, which offers different encodings for characters that look the same (Moran and Cysouw 2017: 54). While scholars working qualitatively will have no problems to see that ‹ə› (Unicode 0259, *Latin Small Letter Schwa*) and ‹ə› (Unicode 01DD, *Latin Small Letter Turned E*) are identical, the two symbols are different for a computer, as they are represented internally by different code points. As a result, an automatic aggregation of data will treat these symbols as different sounds when comparing languages automatically, or when aggregating information on the sound inventories of the languages in the world.

[9] We recognise however, that in some cases it may be more principled to write e.g. /ph/ rather than /p^{h}/. An example is Khmer, where there is good evidence that these aspirated stops are actually clusters, as the /p/ and the /h/ can be separated by an infix (Jakob 1963).

Judging from the above-mentioned examples, we can identify three major problems which make it hard for us to compare phonetic transcriptions cross-linguistically: (a) errors introduced due to the wrong application of the Unicode standard; (b) general incomparability due to the use of different transcription systems; and (c) ambiguities introduced by scholars due to individual transcription preferences. In order to render our transcription systems and datasets cross-linguistically comparable, both for humans and for machines, it therefore seems indispensable to work on a system that normalises transcriptions across different transcription systems and transcription data by linking existing transcription systems and datasets to a unified standard. Such a system should ideally (a) ease the *process of writing phonetic transcriptions* (e.g. by providing tools that automatically check and normalise transcriptions while scholars are creating them), (b) ease the *comparison of existing transcriptions* (e.g. by providing an internal reference point for a given speech sound which links to different grapheme representations in different transcription systems and datasets), and (c) provide a *standard* against which scholars can test existing data. While such an approach cannot solve the theoretical issues of comparability discussed above, it can nonetheless be of considerable practical benefit.

3. The Framework of Cross-Linguistic Transcription Systems

In the spirit of *reference catalogues* for cross-linguistically linked data (Glottolog and Concepticon, see Section 1), we have established a preliminary version of a reference catalogue for phonetic transcription systems and datasets, called *Cross-Linguistic Transcription Systems* (CLTS). The goal of the CLTS framework is to systematically increase the comparability of linguistic transcriptions by linking graphemes generated by transcription systems and graphemes documented in transcription datasets to unique feature bundles drawn from a simple but efficient feature system. With due respect to all obstacles which the documentation of speech through transcription may face in theory and practice, the CLTS system can be seen as a first step towards identifying graphemes across transcription systems and transcription datasets with unique speech sounds. In this sense, CLTS also aids the *translation* between transcription systems and datasets, and can further serve as a *standard* for transcription in practice. Figure 1 illustrates this integrative role of CLTS.

In the following, we will briefly introduce the basic techniques by which we try to render linguistic transcription data comparable. Apart from the data itself

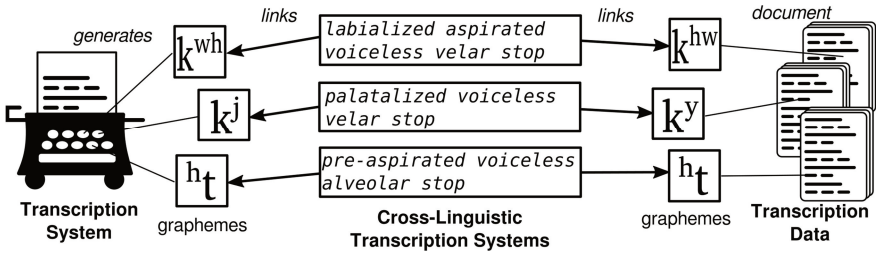

Figure 1. Basic idea behind the CLTS reference catalogue.

(discussed in Section 3.1), which we assemble and annotate in our reference catalogue, we also introduce a couple of different techniques which help to check the consistency of our annotations and ease the creation of new data to which we can link (Section 3.2).

3.1. Materials

3.1.1 Sound classes in CLTS

In order to link graphemes in transcription systems and transcription datasets to feature bundles, it is useful to distinguish rudimentary *classes* of sounds.[10] We distinguish three basic sound classes (*consonants*, *vowels*, and *tones*), a specific class of *markers* (to indicate syllable or morpheme breaks or word boundaries) and two derived classes (*consonant clusters* and *diphthongs*). As of the moment, we do not allow for triphthongs and clusters of more than two consonants (although they could be added at a later stage), in order to keep the system manageable. Clicks are represented as a specific type of consonant that has *click* or *nasal-click* as its *manner*. The representation of tones as a sound class of itself is necessitated by the fact that many phonetic descriptions of tone languages (especially in South-East-Asian languages) represent tone separately. It is further justified by phonological theory, given that tones in many languages may change independently, often correlated with factors that cannot be tied to a seg-

[10] We know that the distinction between basic sound types like *vowels* and *consonants* is often disputed in discussions on phonology and phonetics. For the purpose of linking speech sounds across datasets, however, it is useful to maintain the distinction for practical reasons, as both transcription systems and transcription datasets often maintain these distinctions.

mental context. In addition, we allow tones to be represented with diacritics on vowels (e.g., ‹á› in IPA would be described as an *unrounded open front vowel with high tone*), but we do not encourage scholars to represent their data in this form, as it has many disadvantages when it comes historical language comparison in practice and does not account well for the largely suprasegmental nature of tones.

Complex sound classes in CLTS are not explicitly defined, but instead *automatically derived* by identifying the basic graphemes of which they consist. Diphthongs are thus defined by two vowels, and the grapheme ‹oe›, for example, is treated as a diphthong consisting of a *rounded close-mid back* and an *unrounded close-mid front* vowel. In a similar way, we allow complex consonant clusters to be defined in order to transcribe, for example, doubly articulated consonants or clicks containing a pulmonic release (see Table 1 for examples).[11]

Table 1. Examples for the basic classes of sounds represented in CLTS.

Class	Grapheme	Features
consonant	k^{wh}	labialised aspirated velar stop
vowel	u̯	creaky rounded close back
cluster	kp	from voiceless velar stop to voiceless bilabial stop
diphthong	au̯	from unrounded open front to non-syllabic rounded close back
tone	²¹⁴	contour from-mid-low via-low to-mid-high
marker	+	marker for morpheme boundaries

3.1.2. Features bundles as comparative concepts

In order to ensure that we can compare sounds across different transcription systems and datasets, a feature system that can be used to model sounds as feature bundles, serving as comparative concepts in the sense of Haspelmath (2010) is

[11] For clusters involving clicks, we follow Traill (1993), Güldemann (2001), and Nakagawa (2006), who identify two segments for these sounds, a lingual influx (consonant-onset), and a pulmonic efflux (consonant-offset). For example, [ǀχ] is analyzed as a cluster consisting of a dental click [ǀ] as C-onset, and a uvular fricative [χ] as C-offset.

indispensable. We therefore propose specific feature systems for each of our three sound classes (consonant, vowel, tone), which allow us to identify a large number of different sounds across transcription systems and transcription datasets. The features themselves can be roughly divided into *obligatory features* (like *manner*, *place*, and *phonation* in consonants, and *roundedness*, *height*, and *centrality* in vowels), and *optional features* (usually binary, i.e., present or absent, such as *duration*, *nasalisation*, *aspiration*). Our current feature system contains 25 consonant features,[12] 21 vowel features,[13] and 4 tonal features[14] (Appendix A gives a table with all features and their possible values).

Our choice of features derives from the graphemic representation of sounds in the system of the IPA. It is practically oriented and does not claim to represent any deeper truth about distinctive features in phonology. Instead we focus on being able to align the features as easily as possible with a given graphemic representation of a particular sound in a transcription system. As a result, some features may appear awkward and even naïve from a phonological perspective. For example, instead of distinguishing *ejectives* from plain consonants by manner only (contrasting "ejective stops" and "plain stops"), we code *ejectivity* as an additional feature with a binary value (present or absent). In a similar way, we do not distinguish between different *kinds* of phonation (*voiced*, *breathy-voiced*, *creaky-voiced*, etc.) but code separately for *breathiness*, *creakiness*, and *phonation* (*voiced* or *voiceless*). The advantage of this coding practice is that we can easily infer sounds that we have not yet listed in our database based on the combination of base graphemes and diacritics. In addition, we can also avoid discussions in those cases where linguists often disagree. If we explicitly treated the diacritic ‹ʱ› in the IPA transcription system as indicating breathiness and implying voiced phonation, we would have a problem in distinguishing the admittedly rare instances where scholars explicitly transcribe voiceless stops with breathy release using a voiceless stop in combination with the diacritic for breathy voice (‹pʱ›, ‹tʱ›, ‹kʱ›, etc.) in order to indicate a voiceless initial with

[12] The features are: articulation, aspiration, breathiness, creakiness, duration, ejection, glottalisation, labialisation, laminality, laterality, *manner, nasalisation, palatalisation, pharyngealisation, *phonation, *place, preceding, raising, relative articulation, release, sibilancy, stress, syllabicity, velarisation, and voicing (features with an asterisk are obligatory).

[13] These are: articulation, breathiness, *centrality, creakiness, duration, frication, glottalisation, *height, nasalisation, pharyngealisation, raising, relative articulation, rhotacisation, *roundedness, rounding, stress, syllabicity, tone, tongue root, velarisation, voicing (features with an asterisk are obligatory).

[14] Tonal features are: contour, end, middle, and start (all obligatory).

(breathy) voiced aspiration (Starostin 2017). We could of course argue that these pronunciations are impossible physiologically and impose a system that automatically normalises these graphemes by either treating them as breathy-voiced stops or by treating them as plain-aspirated stops. We prefer, however, to leave the system as inclusive as possible for the time being, following the general principle that it is easier to reduce a given system at a later point for a specific purpose (while preserving the more complex version) than to impose restrictions too early. Given the flexibility of our system (which is presented in more detail in Section 3.2), it would be straightforward to create a strict feature representation that normalises those segments articulatory phoneticians consider impossible. However, if we erroneously reduce the data now, based on assumptions about phonetics that may well be disputed among experts, we run the risk of making regrettable decisions that are difficult to reverse. For this reason, we describe the grapheme ‹pʱ› as a *breathy voiceless bilabial stop consonant*, knowing well that scholars might object to the existence of this sound.

3.1.3. Transcription systems

A transcription system is understood as a *generative entity* in CLTS, being capable of creating sounds that were not produced explicitly before (although the ultimate productivity of a transcription system is, of course, limited). Transcription systems are defined by providing graphemes for the basic sound classes (*consonants*, *vowels*, *tones*), which are explicitly defined and linked to our feature system. Additionally, *diacritics* can be defined and may precede or follow the base graphemes, adding one additional feature per symbol to the base grapheme, depending on their position and the sound class of the base grapheme. In the IPA system, for example, the diacritic ‹ʰ› can only be attached to consonants, but it will evoke different feature values when preceding ‹ʰt› (*pre-aspirated voiceless alveolar stop consonant*) or following ‹tʰ› (*aspirated voiceless alveolar stop consonant*) the base grapheme ‹t›.

Transcription systems can furthermore specify *aliases*, both for base graphemes and for diacritics. The IPA, for example, allows one to indicate *breathiness* by two diacritics, the ‹dʱ› which we mentioned above, and the ‹◌̤›, which is placed under the base grapheme. In the CLTS framework, both glyphs can be parsed, and both ‹dʱ› and ‹d̤› would be interpreted as a *breathy voiced alveolar stop*, but ‹dʱ› would be treated as the regular grapheme representation and ‹d̤› as

its alias.[15] Other important examples of aliases are affricates such as the *voiceless alveolar affricate*, which can be rendered as either a single symbol ‹ʦ› (Unicode 02A6) or two symbols ‹ts› (Unicode points 0074 and 0073, the preferred version in CLTS).[16] In these and many other cases, the CLTS framework correctly recognises the sounds denoted by the graphemes, while at the same time proposing a default representation of ambiguous graphemes in a given transcription system.

CLTS currently offers five different transcription systems, namely a *broad* version of the IPA (called BIPA), a preliminary version of the transcription system underlying the *Global Lexicostatistical Database* (GLD, http://starling.rinet.ru/new100/main.htm, Starostin and Krylov 2011), the transcription system employed by the *Automatic Similarity Judgment Project* (ASJPCODE, http://asjp.clld.org, Wichmann et al. 2016), an initial version of the *North American Phonetic Alphabet* (NAPA, Pullum and Ladusaw 1996), and an initial version of the Uralic Phonetic Alphabet (UPA, Setälä 1901). Most of our initial efforts went into the creation of the B(road)IPA system. This choice is justified, as most transcription datasets also follow the supposed IPA standards to a large degree. In the future, however, we hope that we can further expand the data by expanding both the generative power and the accuracy of the remaining transcription systems, and by adding new transcription systems.

3.1.4. Transcription data

CLTS currently links 15 different transcription datasets, summarised in Table 2. The datasets were selected for different reasons. We tried to assemble as many of the cross-linguistic sound inventory datasets as possible (Nikolaev 2015; Maddieson et al. 2013; Mielke 2008; Moran et al. 2014; Ruhlen 2008), since apart from the comparison of Phoible with Ruhlen's database by Dediu and Moisik (2016), these existing datasets have not yet been thoroughly compared. Linking them to CLTS should thus immediately illustrate the usefulness of our

[15] The decision of what we define as an alias and what we define as the regular symbol is mostly based on practical considerations regarding visibility. Since the glyph ‹◌̤› will be difficult if not impossible to spot when placed under certain consonants, we decided to define ‹ʱ› as the base diacritic to indicate breathiness for consonants, but kept ‹◌̤› for vowels.

[16] We know well that no single decision will ever satisfy all users, but given the flexibility of the system, users can always easily define their sub-standard while at the same time maintaining comparability via our feature system.

Table 2. Basic coverage statistics for transcription datasets linked by the CLTS framework.

ID	Name	Source	Graph.	CLTS	Cov.
APiCS	Atlas of Pidgin and Creole Language Structures Online	Michaelis et al. 2013	177	177	100
BDPA	Benchmark database of phonetic alignments	List and Prokić 2014	1466	1329	91
BJDX	Chinese Dialect Vocabularies	Beijing Daxue 1964	124	124	100
Chomsky	Sound Pattern of English	Chomsky and Halle 1968	45	45	100
Diachronica	Index Diachronica	Anonymous 2014, D. 2017	652	552	85
Eurasian	Database of Eurasian Phonological Inventories	Nikolaev 2015	1562	1366	87
LAPSyD	Lyon-Albuquerque Phonological Systems Database	Maddieson et al. 2013	795	712	90
Multimedia	Multimedia IPA Charts	Department of Linguistics 2017	138	134	97
Nidaba	Lexicon Analysis and Comparison	Eden 2018	1936	1872	97
PanPhon	PanPhon Project	Mortensen 2017	6334	6220	98
PBase	PBase Project	Mielke 2008	1068	859	80
Phoible	Phonetics Information Base and Lexicon	Moran et al. 2014	1843	1589	86
PoWoCo	Potential of Word Comparison	List et al. 2017	378	370	98
Ruhlen	Global Linguistic Database	Ruhlen 2008	701	437	62
Wiki	Wikipedia IPA Descriptions	Wikipedia contributors 2017	184	168	91

framework (see Section 4.3 for details). Furthermore, given the large number of sound segments which one can find in these datasets (most of them representing

a supposedly strict version of IPA), they provide a useful way to test how well our framework recognises sounds written in IPA which were not explicitly defined. Additional datasets were chosen to illustrate links to feature systems (Chomsky and Halle 1968), for illustrative purposes (Department of Linguistics 2017; Wikipedia contributors 2018), or to test our system by providing either large collections of graphemes (Eden 2018; Mortensen 2017; List and Prokić 2014; List et al. 2017), or for reasons of general interest and curiosity (Michaelis et al. 2013; Anonymous 2014).

Table 3. Small excerpt of Unicode confusables normalised in CLTS.

Source	Code	Target	Code	Sound Name
λ	03BB	ʎ	028E	palatalised alveolar lateral approximant consonant
ə	01DD	ə	0259	unrounded mid central vowel
ʔ	0242	ʔ	0294	voiceless glottal stop consonant
ε	03B5	ɛ	025B	unrounded open-mid front

3.2. Methods

3.2.1. Parsing and generating sounds

CLTS employs a sophisticated algorithm for the parsing and generation of graphemes for a given transcription system. The parsing algorithm employs a three-step procedure, consisting of (A) normalisation, (B) direct lookup, and (C) generation of graphemes.

In (A), all sounds are generally normalised, following Unicode's NFD normalisation in which diacritics and base graphemes are maximally dissolved (Moran and Cysouw 2017: 16). In addition, the algorithm uses system-specific normalisation tables of homoglyphs, which can be easily confused. The normalisation applies to single glyphs only and employs a simple lookup table in which source and target glyph are defined. In this way, one can easily prevent users from using the wrong character to represent, for example, the schwa-sound [ə], since the data is normalised beforehand. Table 3 gives a small list of examples for base graphemes normalised in CLTS.

In (B), the algorithm searches for direct matches of the grapheme with the base graphemes provided along with the transcription system. If a grapheme can be matched directly, the algorithm checks whether it is flagged as an alias and provides the corrected grapheme.

If the grapheme could not be resolved in (A), the algorithm tries to generate it in (C), by first using a regular expression to identify whether the unknown grapheme contains a known base grapheme. If this is the case, the algorithm searches to the left and the right of the base grapheme for known diacritics, looks up the features from the table of diacritic features, and then combines the features of the base grapheme with the new features supplied by the diacritics to a generated sound. The algorithm returns an unknown sound if either no base grapheme can be identified or if one of the diacritics cannot be interpreted correctly.[17]

The algorithm can be used in a reverse fashion by supplying a feature bundle from which the algorithm will then try to infer the underlying grapheme in a given transcription system. Here again, we can distinguish between sounds that were already defined as base graphemes of the transcription system, and sounds that are generated by identifying a base sound and then converting the remaining features to diacritic symbols. Since the order of features serving as diacritics is defined directly, the algorithm explicitly normalises phonetic transcriptions in those cases in which features are supplied in the wrong order. For example, if a transcription system provides the *labialised aspirated voiceless velar stop consonant* as ‹k^{hw}› (as, for example, APiCS), the algorithm will normalise the order of diacritics to ‹k^{wh}› and flag the grapheme as an alias.

3.2.2. Python API and online database

CLTS comes with a Python API which can be used from the command line or within Python scripts and offers a convenient way to test the framework both on large datasets and on an ad-hoc basis. It also comes along with a brief tutorial introducing the main aspects of the code as well as a “cookbook” containing a series of coding recipes to address specific tasks. The data is further presented

[17] The generation procedure is strictly *accumulative*, and no features of the base grapheme can be changed post-hoc. This explains most peculiarities of our feature system and reflects a deliberate design choice. Given the large number of speech sounds that we could identify in the different transcription datasets, we had to make sure to keep the complexity of the algorithm on a level that can still be easily understood.

online at https://clts.clld.org in the form of a database in the well-known *Cross-Linguistically Linked Data* framework (http://clld.org, Haspelmath and Forkel 2015), which provides interested users with the common look and feel of popular CLLD datasets such as Glottolog or WALS. There is also a web application, available at http://calc.digling.org/clts/, that allows users to quickly check if their data conforms to the standards defined in our database. More information on the Python API can be found in Appendices B. The full source code is available online at https://zenodo.org/record/1623511.

4. Examples

4.1. Normalisation and parsing of sounds

In order to illustrate how the parsing algorithm underlying CLTS works, let us consider the grapheme ‹^{w}t^{s}:^{h}› as a fictitious example which we want to parse with the B(road)IPA system of CLTS. In a first step, the algorithm normalises the grapheme, thereby replacing the normal colon ‹:› by its correct IPA equivalent ‹ː›. The colon is often confused with the correct IPA counterpart, and often we find both the colon and the correct glyph in the same dataset (e.g., in APiCS). The remaining sequence ‹^{w}t^{s}ː^{h}› is now tested for direct matches with the table of pre-defined base graphemes of BIPA. Since the algorithm does not find the sequence, it will apply a regular expression to check against potential base grapheme candidates and select the longest grapheme. In our case, this is the sequence ‹t^{s}› which itself is flagged as an alias whose correct version is ‹ts›. In terms of features, this sound is defined as a *voiceless alveolar sibilant affricate consonant*. Two subsequences are remaining, the ‹^{w}› to the left, and ‹ː^{h}› to the right. The first can be directly mapped to the feature value *pre-labialised*, the second subsequence maps to *long* and *aspirated*, respectively. The algorithm now assembles all features to a feature bundle and sorts them according to the pre-defined order of features when writing a grapheme. The resulting sound is now described as a *pre-labialised aspirated long voiceless alveolar sibilant affricate consonant* and the grapheme representation in BIPA is given as ‹^{w}ts^{h}ː›. The sound will be labeled as both *normalised* and *aliased*, accounting for the correction of the homoglyph ‹:›, the alias ‹t^{s}›, and the order of the original grapheme.

Table 4: Parsing examples for the CLTS algorithm.

Input	Norm.	Alias	Base	BIPA	Name
a:	: → :	–	–	a:	long unrounded open front vowel
t:s	: → :	t:s → ts:	–	ts:	long voiceless alveolar sibilant affricate consonant
k^{hw}	–	–	k	k^{wh}	labialised aspirated voiceless velar stop consonant
t^{hy}	^{y} → j		t	t^{jh}	palatalised aspirated voiceless alveolar stop consonant
$t^{:sh}$	–	–	t	?	unknown sound (‹^{s}› is not defined as a diacritic)

Table 4 gives more illustrations of the algorithm by showing the different stages of normalisation, alias lookup, identification of the base grapheme, and generation of the target sound. The last sound in the table cannot be parsed with the current transcription system, since the diacritic ‹^{s}› in the grapheme ‹$t^{:sh}$› is not defined as a valid diacritic (as its interpretation would be ambiguous, since in many transcription systems it is only used in combination with alveolars and dentals to indicate an affricate).

4.2. Looking at transcription datasets through CLTS

Table 2 above provides some general statistics regarding the number of graphemes which we find in the original transcription data, the number of items we could link to CLTS, and the number of unique sounds which we identify. The general statistics reveal a rather disappointing situation: instead of providing largely similar collections of graphemes for the speech sounds collected in the different transcription datasets, we find that only a small proportion effectively overlaps, blowing the number of supposedly unique sounds up to as many as 8754. While this might point to errors in our system, we are confident that it instead displays the general nature of linguistic transcription data, given that the 17403 graphemes of all transcription datasets themselves amount to 12384 unique graphemes *without* CLTS. We further checked the majority of the graphemes manually, finding that it is not the failure of the framework to merge sounds for which spelling variants exist, but rather the fact that many datasets list large numbers of sounds one might judge to be unlikely to be produced in any language and which are of low frequency in their respective datasets. These might well reflect idiosyncrasies of interpretation rather than real variation.

A further factor contributing to the large number of sounds in CLTS are transcription datasets like Nidaba and PanPhon which were at least in part automatically created in order to allow one to recognise and provide features for sounds which were not yet accounted for in the data. Since the CLTS framework has a strong generative component, linking these datasets to our framework is useful for two reasons. First, it allows us to generate a large number of potential sounds that might have already been used in some datasets we have not yet included and will help scholars in linking their data to CLTS. Second, it offers a test for the generative strength of our system. Since CLTS so far creates many more potential sounds, which can be uniquely identified, this is an important proof of concept that our system is already capable of integrating many different transcription datasets in an almost completely automated manner.

What we can also learn from linking transcription data to CLTS are obvious errors in the original datasets. Many datasets, for example, provide different graphemes for what CLTS assigns to the same sound. Examples are ‹ts› vs. ‹tˢ› for the voiceless alveolar sibilant affricate consonant in the Eurasian dataset, since ‹tˢ› only occurs one time in the data, and is assigned to Danish, where it reflects phonological convention rather than real pronunciation. Many datasets also confuse the order of diacritics, thus listing ‹kʰʷ› and ‹kʷʰ› as two separate sounds (Phoible, LAPSyD, Diachronica). Other datasets distinguish ‹t̪s̪› from ‹ʈʂ› (Eurasian, PoWoCo, PBase), of which the latter is defined as alias in the B(road)IPA of CLTS and thus described as *voiceless retroflex sibilant affricate consonant*. Since CLTS normalises the order of diacritics, and provides a large alias system for the BIPA transcription system, these errors can be easily detected and help to improve future versions of the respective datasets.

5. Outlook

Given the theoretical difficulties inherent in phonetic transcription (elaborated in Section 2.3), readers may ask themselves whether linguistics really needs a reference catalogue such as the one we present here. Apart from the immediate benefit of increasing the comparability of large transcription datasets, which we have illustrated above, we see many interesting use-cases for our framework. Given the various methods for normalisation that CLTS offers, the framework can help scholars working with transcriptions to improve their data considerably. This does not only apply to the large phoneme inventory datasets, which

can directly profit from the problems which were identified when linking them to CLTS, but also to the increasing numbers of digitally available lexical datasets resulting from retro-digitisation of older sources or recent field work. With a growing interest in computer-assisted applications in historical linguistics and lexical typology, especially in automated methods for the identification of cognate words (List et al. 2017; Jäger et al. 2017), there is also an increased need for high-quality transcriptions that can be easily parsed by algorithms. With its inbuilt feature system and the feature systems supplied as metadata with the transcription datasets, providing coverage for a large number of sounds, advanced methods for cognate detection and linguistic reconstruction can be easily designed and tested. Last but not least, CLTS also has an educational component, since it rigorously exposes variation across transcription datasets, bringing the need for consistency and adherence to standards to our attention.

References

Anonymous. 2014. *Index Diachronica*. <https://chridd.nfshost.com/diachronica/>

Bell, A. 1867. *Visible speech: The science of universal alphabetics: Or, self-interpreting physiological letters, for the writing of all languages in one alphabet. Illustrated by tables, diagrams, and examples*. London: Simpkin, Marshall.

Běijīng Dàxué 北京大学. 1964. *Hànyǔ fāngyán cíhuì* [Chinese dialect vocabularies]. Běijīng: Wénzì Gǎigé 文字改革.

Bybee, J. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Chao, Y. 2006. A system of 'tone letters'. In: Wu, Z.-J. and X.-N. Zhao (eds.), *Linguistic essays by Yuenren Chao*. Běijīng: Shāngwù. 98–102.

Charpentier, J.-M. and A. François. 2015. *Linguistic atlas of French Polynesia / Atlas linguistique de la Polynésie française*. Berlin, Boston: De Gruyter Mouton.

Chomsky, N. and M. Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Crowley, T. 2006. *The Avava Language of Central Malakula (Vanuatu)*. The Australian National University: Pacific Linguistics, Research School of Pacific and Asian Studies.

Crowley, T. 2006. *Nese: A Diminishing Speech Variety of Northwest Malakula (Vanuatu)*. The Australian National University: Pacific Linguistics, Research School of Pacific and Asian Studies.

Dediu, D. and S. Moisik. 2016. Defining and counting phonological classes in cross-linguistic segment databases. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation*. 1955–1962.

Dench, A. 2002. Descent and diffusion: The complexity of the Pilbara situation. In: Aikhenvald, A. and R. Dixon (eds.), *Areal diffusion and genetic inheritance: Problems in comparative linguistics*. Oxford: Oxford University Press. 105–133.
Dodd, R. 2014. V'ënen Taut: Grammatical topics in the Big Nambas Language of Malekula. (PhD dissertation, University of Waikato.)
Dolgopolsky, A. 1964. Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concering the oldest relationships among the language families of Northern Eurasia]. *Voprosy Jazykoznanija* 2. 53–63.
Dryer, M. and M. Haspelmath. 2011. *The World Atlas of Language Structures online*. Munich: Max Planck Digital Library.
Eden, E. 2018. Measuring phonological distance between languages. (PhD dissertation, University College London.)
Güldemann, T. 2001. Phonological regularities of consonant systems across Khoesan lineages. *University of Leipzig Papers on Africa* 16. 1–50.
Güldemann, T. 2014. 'Khoisan' linguistic classification today. In: Güldemann, T. and A.-M. Fehn (eds.), *Beyond 'Khoisan'. Historical relations in the Kalahari Basin*. Amsterdam and Philadelphia: John Benjamins. 1–40.
Hammarström, H., R. Forkel, and M. Haspelmath. 2017. Glottolog. Version 3.0. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Haspelmath, M. 2010. Comparative concepts and descriptive categories. *Language* 86(3). 663–687.
Haspelmath, M. and R. Forkel. 2015. CLLD – Cross-Linguistic Linked Data. Max Planck Institute for Evolutionary Anthropology: Leipzig.
Herzog, G., S. Newman, E. Sapir, M. Swadesh, M. Swadesh, and C. Voegelin. 1934. Some orthographic recommendations. *American Anthropologist* 36(4). 629–631.
Honeybone P. 2005. Diachronic evidence in segmental phonology: The case of laryngeal specifications. In: van Oostendorp, M. and J. van de Weijer (eds.), *The internal organisation of phonological segments*. Mouton de Gruyter: Berlin and New York. 319–354.
Hóu Jīngyī 侯精一 (ed.). 2004. *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库 [Phonological database of Chinese dialects]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.
Huáng, B. and X. Liào. 2002. *Xiàndài Hànyǔ* 现代汉语 [Modern Chinese]. Běijīng: Gāoděng Jiàoyù.
International Institute of African Languages and Cultures. 1930. *Practical orthography of African languages*. (Revised edition.) Oxford: Oxford University Press.
International Phonetic Association. 1912. *The Principles of the International Phonetic Association*. Bourg-la-Reine and London: Paul Passy and Daniel Jones.
International Phonetic Association. 1999. *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.
International Phonetic Association. 2015. *The International Phonetic Alphabet*. (Revised to 2015.)
Department of Linguistics. 2017. *Multimedia IPA chart*. Victoria: University of Victoria.

Jacob, J.M. 1963. Prefixation and infixation in old Mon, old Khmer, and modern Khmer. *Linguistic comparison in Southeast Asia and the Pacific*. 62–70.

Jäger, G., J.-M. List and P. Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. (Long papers.) 1204–1215.

Kalusky, W. 2017. *Die Transkription der Sprachlaute des Internationalen Phonetischen Alphabets: Vorschläge zu einer Revision der systematischen Darstellung der IPA-Tabelle*. München: LINCOM Europa.

Kieviet, P. 2017. *A Grammar of Rapa Nui*. Berlin: Language Science Press.

Köhler, O., P. Ladefoged, J. Snyman, A. Traill and R. Vossen. 1988. The symbols for clicks. *Journal of the International Phonetic Association* 18(2). 140–142.

Kümmel, M. 2008. *Konsonantenwandel* [Consonant change]. Reichert: Wiesbaden.

Lepsius, C. 1854. *Das allgemeine linguistische Alphabet: Grundsätze der Übertragung fremder Schriftsysteme und bisher noch ungeschriebener Sprachen in europäische Buchstaben*. Wilhelm Hertz: Berlin.

List, J.-M. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

List, J.-M. and J. Prokić. 2014. A benchmark database of phonetic alignments in historical linguistics and dialectology. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. 288–294.

List, J.-M., M. Cysouw, and R. Forkel. 2016. Concepticon. A resource for the linking of concept lists. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. 2393–2400.

List, J.-M., S. Greenhill, and R. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1). 1–18.

Lynch, J. 2016. Malakula internal subgrouping: Phonological evidence. *Oceanic Linguistics* 55(2). 399–431.

Maddieson, I. 1984. *Patterns of sounds*. Cambridge: Cambridge University Press.

Maddieson, I., S. Flavier, E. Marsico, C. Coupé and F. Pellegrino. 2013. LAPSyD: Lyon-Albuquerque Phonological Systems Database. In: *Proceedings of Interspeech*.

Malau, C. 2016. *A grammar of Vurës, Vanuatu*. Berlin: Walter de Gruyter:

Mann, M. and D. Dalby. 1987. *A thesaurus of African languages: A classified and annotated inventory of the spoken languages of Africa with an appendix on their written representation*. London: Zell Publishers.

Michaelis, S., P. Maurer, M. Haspelmath and M. Huber. 2013. *The Atlas of Pidign and Creole language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Mielke, J. 2008. *The emergence of distinctive features*. Oxford: Oxford University Press.

Moran, S. and M. Cysouw. 2017. *The Unicode cookbook for linguists. Managing writing systems using Orthography Profiles*. Zürich: Zenodo.

Moran, S., D. McCloy and R. Wright (eds.). 2014. *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Mortensen, D. 2017. *PanPhon. Python API for accessing phonological features of IPA Segments*. Pittsburgh: Carnegie Mellon School of Computer Science.

Nakagawa, H. 2006. Aspects of the phonetic and phonological structure of the Gui language. (PhD dissertation, University of the Witwatersrand, Johannesburg.)
Nikolaev, D., A. Nikulin and A. Kukhto. 2015. *The database of Eurasian phonological inventories*. Moscow: RGGU. <http://eurasianphonology.info>
Press, M. L. 1980. *Chemehuevi: A grammar and lexicon*. Berkeley: University of California Press.
Pullum, G. and W. Ladusaw. 1996. *Phonetic symbol guide*. Chicago: University of Chicago Press.
Ruhlen, M. 2008. *A global linguistic database*. Moscow: RGGU.
Salisbury, M.C. 2002. A grammar of Pukapukan. (PhD dissertation, The University of Auckland.)
Sapir, E. 1930. *Southern Paiute, a Shoshonean language*. Boston: Academic Press.
Saussure, F. de. 1878. *Mémoire sur le système primitif des voyelles dans les langues indo-européennes*. Leipzig: Teubner.
Saussure, F. de. 1916. *Cours de linguistique générale*. Lausanne: Payot.
Setälä, E. 1901. Über transskription der finnisch-ugrischen sprachen. *Finnisch-ugrische Forschungen* 1. 15–52.
Simpson, A. 1999. Fundamental problems in comparative phonetics and phonology: does UPSID help to solve them. In: *Proceedings of the 14th international congress of phonetic sciences*.
Starostin, G. and P. Krylov (eds.). 2011. *The global lexicostatistical database. Compiling, clarifying, connecting basic vocabulary around the world: From free-form to tree-form*. <http://starling.rinet.ru/new100/main.htm>
Starostin, G. (ed.) 2017. *Annotated Swadesh wordlists for the Hmong group* (Hmong-Mien family).
Stimson, J. F. and D.S. Marshall. 1964. *A dictionary of some Tuamotuan dialects of the Polynesian language*. Leiden: M. Nijhoff.
Sweet, H. 1877. *A handbook of phonetics, including a popular exposition of the principles of spelling reform*. Oxford: Clarendon Press.
Tadadjeu, M. and E. Sadembouo. 1979. *Alphabet Générale des langues Camerounaises*. Yaoundé: Departement des Langues Africaines et Linguistique, Université de Yaoundé.
Traill A. 1993. The feature geometry of clicks. In: van Staden, P.M.S. (ed.), *Linguistica: Festschrift E. B. van Wyk: 'n huldeblyk*. Pretoria: van Schaik. 134–140.
Tregear, E. 1899. *Dictionary of Mangareva: Or Gambier Islands*. Wellington: J. Mackay.
Trubetzkoy, N. 1939. *Grundzüge der Phonologie* [Foundations of phonology]. Prague: Cercle Linguistique de Copenhague.
UNESCO. 1978. African languages. In: *Proceedings of the meeting of experts on the transcription and harmonization of African languages*.
Wichmann, S., E. Holman and C. Brown. 2016. *The ASJP database*. Jena: Max Planck Institute for the Science of Human History.
Wikipedia contributors. 2018. International Phonetic Alphabet. Wikipedia, The Free Encyclopedia. <https://en.wikipedia.org/w/index.php?title=International_Phonetic_Alphabet&oldid=822828531>. Accessed 29 Jan 2018.

Acknowledgments

JML and TT were funded by the the ERC Starting Grant 715618 "Computer-Assisted Language Comparison" (http://calc.digling.org). We thank Gereon Kaiping for providing early support in testing and discussing the *pyclts* software package. We thank Adrian Simpson, Martin Haspelmath, Ludger Paschen, and Paul Heggarty for helpful comments on earlier versions of this draft, and we thank Simon J. Greenhill and Christoph Rzymski for providing support with the software.

Software and data

Software and data accompanying this paper have been hosted with Zenodo and can be found at https://doi.org/10.5281/zenodo.1617697. The source code for the Python API is curated on GitHub at https://github.com/cldf/clts/. The data can be further inspected and conveniently browsed at https://clts.clld.org.

Appendix

Current feature system underlying the CLTS framework.

Sound type	**Feature**	**Value**
vowel	relative_articulation	centralized
vowel	relative_articulation	mid-centralized
vowel	relative_articulation	advanced
vowel	relative_articulation	retracted
vowel	centrality	back
vowel	centrality	central
vowel	centrality	front
vowel	centrality	near-back
vowel	centrality	near-front

Sound type	Feature	Value
vowel	creakiness	creaky
vowel	rounding	less-rounded
vowel	rounding	more-rounded
vowel	stress	primary-stress
vowel	stress	secondary-stress
vowel	pharyngealization	pharyngealized
vowel	rhotacization	rhotacized
vowel	voicing	devoiced
vowel	nasalization	nasalized
vowel	syllabicity	non-syllabic
vowel	raising	lowered
vowel	raising	raised
vowel	height	close
vowel	height	close-mid
vowel	height	mid
vowel	height	near-close
vowel	height	near-open
vowel	height	open
vowel	height	open-mid
vowel	frication	with-frication
vowel	roundedness	rounded
vowel	roundedness	unrounded
vowel	duration	long
vowel	duration	mid-long
vowel	duration	ultra-long
vowel	duration	ultra-short

Sound type	Feature	Value
vowel	velarization	velarized
vowel	tongue_root	advanced-tongue-root
vowel	tongue_root	retracted-tongue-root
vowel	tone	with_downstep
vowel	tone	with_extra-high_tone
vowel	tone	with_extra-low_tone
vowel	tone	with_falling_tone
vowel	tone	with_global_fall
vowel	tone	with_global_rise
vowel	tone	with_high_tone
vowel	tone	with_low_tone
vowel	tone	with_mid_tone
vowel	tone	with_rising_tone
vowel	tone	with_upstep
vowel	articulation	strong
vowel	breathiness	breathy
vowel	glottalization	glottalized
consonant	aspiration	aspirated
consonant	sibilancy	sibilant
consonant	creakiness	creaky
consonant	release	unreleased
consonant	release	with-lateral-release
consonant	release	with-mid-central-vowel-release
consonant	release	with-nasal-release
consonant	ejection	ejective
consonant	place	alveolar

Sound type	Feature	Value
consonant	place	alveolo-palatal
consonant	place	bilabial
consonant	place	dental
consonant	place	epiglottal
consonant	place	glottal
consonant	place	labial
consonant	place	linguolabial
consonant	place	labio-palatal
consonant	place	labio-velar
consonant	place	labio-dental
consonant	place	palatal
consonant	place	palatal-velar
consonant	place	pharyngeal
consonant	place	post-alveolar
consonant	place	retroflex
consonant	place	uvular
consonant	place	velar
consonant	pharyngealization	pharyngealized
consonant	voicing	devoiced
consonant	voicing	revoiced
consonant	nasalization	nasalized
consonant	preceding	pre-aspirated
consonant	preceding	pre-glottalized
consonant	preceding	pre-labialized
consonant	preceding	pre-nasalized
consonant	preceding	pre-palatalized

Sound type	Feature	Value
consonant	labialization	labialized
consonant	syllabicity	syllabic
consonant	palatalization	labio-palatalized
consonant	palatalization	palatalized
consonant	phonation	voiced
consonant	phonation	voiceless
consonant	duration	long
consonant	duration	mid-long
consonant	stress	primary-stress
consonant	stress	primary-stress
consonant	stress	primary-stress
consonant	stress	primary-stress
consonant	stress	secondary-stress
consonant	laterality	lateral
consonant	velarization	velarized
consonant	manner	affricate
consonant	manner	approximant
consonant	manner	click
consonant	manner	fricative
consonant	manner	implosive
consonant	manner	nasal
consonant	manner	nasal-click
consonant	manner	stop
consonant	manner	tap
consonant	manner	trill
consonant	laminality	apical

Sound type	Feature	Value
consonant	laminality	laminal
consonant	articulation	strong
consonant	breathiness	breathy
consonant	glottalization	glottalized
consonant	raising	lowered
consonant	raising	raised
consonant	relative_articulation	centralized
consonant	relative_articulation	mid-centralized
consonant	relative_articulation	advanced
consonant	relative_articulation	retracted
tone	middle	via-high
tone	middle	via-low
tone	middle	via-mid
tone	middle	via-mid-high
tone	middle	via-mid-low
tone	start	from-high
tone	start	from-low
tone	start	from-mid
tone	start	from-mid-high
tone	start	from-mid-low
tone	start	neutral
tone	contour	contour
tone	contour	falling
tone	contour	flat
tone	contour	rising
tone	contour	short

Sound type	Feature	Value
tone	end	to-high
tone	end	to-low
tone	end	to-mid
tone	end	to-mid-high
tone	end	to-mid-low

Corresponding author:
Johann-Mattis List
Department of Linguistic and Cultural Evolution
Max Planck Institute for the Science of Human History
Kahlaische Straße 10
07745 Jena
Germany
list@shh.mpg.de