SOUND COMPARISONS: A NEW ONLINE DATABASE AND RESOURCE FOR RESEARCH IN PHONETIC DIVERSITY

Paul Heggarty *et al.* — *A full list of authors, affiliations and emails appears at the end of the paper*

Max Planck Institute for the Science of Human History — *et al.*
Paul.Heggarty@gmail.com — *et al.*

ABSTRACT

Sound Comparisons hosts over 90,000 individual word recordings and 50,000 narrow phonetic transcriptions from 600 language varieties from eleven language families around the world. This resource is designed to serve researchers in phonetics, phonology and related fields. Transcriptions follow new initiatives for standardisation in usage of the IPA and Unicode. At soundcomparisons.com, users can explore the transcription datasets by phonetically-informed search and filtering, customise selections of languages and words, download any targeted data subset (sound files and transcriptions) and cite it through a custom URL. We present sample research applications based on our extensive coverage of regional and sociolinguistic variation within major languages, and also of endangered languages, for which Sound Comparisons provides a rapid first documentation of their diversity in phonetics. The multilingual interface and user-friendly, 'hover-to-hear' maps likewise constitute an outreach tool, where speakers can instantaneously hear and compare the phonetic diversity and relationships of their native languages.

Keywords: database, endangered languages, sound change, comparative/historical linguistics, diversity.

1. PURPOSE: A RESOURCE FOR PHONETIC RESEARCH

Sound Comparisons is a major collaborative project that since 2002 has collected over 90,000 individual word recordings from 600 language varieties of eleven language families around the world. Of these recordings, over 50,000 are accompanied by the corresponding narrow phonetic transcription, conforming to the latest approaches to standardising usage of IPA and Unicode (see §5 below), to ensure their utility also for the growing use of computational analyses in phonetics. The fundamental phonetic objective entails a focus on collecting wordforms that are *cognate* within each language family covered, to ensure the most direct and meaningful comparisons between different realisations of the same original form.

This paper focuses on the potential of the *Sound Comparisons* data-set, from its wide range of linguistic contexts, to support research in phonetics, that can also be applied to other fields (comparative/historical linguistics, dialectology, sociolinguistics). To make the most of this potential, *Sound Comparisons* is founded equally on a powerful online interface to explore its data-set. This offers the functionality needed to turn *Sound Comparisons* into a resource and research tool for phoneticians. It includes powerful functions to search and filter the phonetic diversity represented in the database, to customise the selection and display of data, and to download and cite any such targeted selections (§6). Having presented the database, its coverage and functionality, this paper closes with brief illustrations of some sample research applications.

2. DATABASE STRUCTURE AND COVERAGE

The *Sound Comparisons* data-set is structured as a series of 'studies' (currently ten), by language family and/or by region. As a framework for the collection, storage and exploration of data on phonetic diversity, *Sound Comparisons* can in principle host data on any language family or region worldwide. Actual coverage so far is a function of the specialisms and research objectives of the main researchers involved, and the availability of time and funding for fieldwork to make the recordings.

- In Europe (6 studies): four on Romance, Germanic, Balto-Slavic and Celtic (and parts of their dialectal diaspora worldwide); a Europe study on their overlap in deep Indo-European cognates; and a highly focused study on dialectal variation within English.
- In South America (3): on the Central Andes (the Quechua, Aymara and Uru families); on Mapudungun in Patagonia; and a pilot study on the indigenous languages of Brazil.
- Vanuatu (1): covering 40+ languages and 80 dialectal variants on the linguistically hyper-diverse island of Malakula, and its neighbours.

Studies also include, where known, assumed and reconstructed transcriptions for earlier historical stages of the modern languages recorded, such as Shakespearean English, Old High German, Classical

Latin or Proto-Quechua. These are of particular use for comparative and historical purposes (see §6).

3. DIVERSITY AND DOCUMENTATION

Much of the *Sound Comparisons* coverage thus far has prioritised languages that are highly endangered, indeed moribund. This reflects an urgent effort in language documentation, of a new form that can be considered 'shallow but broad'. For although the data collected for any one language variety are but a limited word-list, those lists are devised to provide at least a representative sample of the phonetics of that variety. Moreover, they can be collected relatively quickly, in the space of a few hours from a suitable native-speaker informant. Before linguists can document them in richer detail, many of the varieties recorded will have vanished — some already have.

A further major objective of *Sound Comparisons* is outreach, to return something to the informants' speaker communities, including contributing to efforts towards awareness, native-language literacy and revitalisation. The explorer website is specifically designed to be as user-friendly as possible to the general public, not least speakers of the languages themselves. It includes a multilingual user-interface, so that the languages of Vanuatu can be explored through the entire website in a Bislama-language version (the English-based creole that is the *lingua franca* of Vanuatu).

For researchers in phonetics, *Sound Comparisons* thus constitutes a mass of detailed phonetic data on hundreds of languages and dialects that are otherwise very little documented, and for which few or no other recordings and resources are easily available — let alone in a form immediately and directly comparable with cognate forms in scores of other related language varieties.

4. DATA SELECTION CRITERIA

Within each study, the data are structured on two axes: language varieties, and words. (Both can be searched and customised, independently: see §6.) 'Language' refers to any lect: e.g. 49 modern and 10 historical varieties of English, 37 of Mapudungun.

Sampling of language varieties has been determined by two main criteria: to be representative of linguistic and dialectal diversity, and urgency in the face of the imminent extinction that hangs over much of that diversity. The complex balance between these criteria often overrides the default of sampling evenly through geographical space. Fuller explanations of the sampling policies followed is given in [1], and a series of subsequent publications in preparation on each individual study.

Examples of urgently targeted languages include the Low German originally native to Pomerania, but now spoken only by a last generation particularly in Wisconsin, Iowa and Brazil (descendants of emigrants in the 1850s). Similar motivations have driven the *Sound Comparisons* coverage of Volga German, Transylvanian 'Saxon' and Pennsylvania 'Dutch', Patagonian Welsh, the Scottish Gaelic of Nova Scotia, Chabacano Spanish 'creole' in the Philippines, both varieties of Sorbian, and so on.

The ten studies in *Sound Comparisons* fall into two different types. Linguists might instinctively assume that the words axis is based on a Swadesh-type reference list of basic comparison *meanings* such as HEAD, LEG or DOG, for which *Sound Comparisons* would collect the word used in each language. But this actually applies only to the latest two studies: the special case of Vanuatu, and the pilot study on Brazil, which do work to modified regional versions of the Swadesh 200-meaning list.

All other studies follow a different rationale, however, in line with the fundamentally phonetic and comparative motivation behind "*sound* comparisons". To that end, it is not meanings but *cognates* that allow for the most valuable, direct and meaningful comparisons of sounds between different languages. Cognates constitute extensive sets of phonetic realisations that may differ slightly or very significantly, but which all correspond to each other as reflexes of the same original underlying form. Such cognate sets can span the diversity across the dialects and languages of an entire language family, as well as numerous varieties (geolectal and sociolectal) of a single language.

This is why each family constitutes a separate study, with its own list of cognates as its 'word' axis. The Germanic study, for instance, is based on c. 100 cognates that have survived in (ideally) all languages and dialects across that family. This is not a list of meanings like HEAD, LEG or DOG, then, but of sets of cognates in Germanic — which may have those meanings in some languages, but not in all. So English *head* is covered alongside its cognate *Haupt* in standard German, rather than its meaning equivalent *Kopf*. (*Haupt* now means 'main', but like English *head*, it goes back to the same Proto-Germanic *xaubadan.) Similarly, *Bein* may mean LEG in standard German, but is set alongside English *bone*, its true cognate (from *bainan). In some Upper German dialects, too, *Bein* means 'bone'.

In the Romance family, though, no single cognate survives well across the family for any of the Latin source terms for HEAD, LEG or DOG. Necessarily, for Romance the *Sound Comparisons* cognate list is different, and based on Latin as the reference for common origin and cognacy. The meanings in

which cognates happen to survive widely vary significantly from one family to the next. The Europe 'overlap' study has just twenty deep Indo-European cognates across the four European studies.

As for how many and which cognates are covered in each study, even within an individual branch like Germanic, Romance or Celtic, diversity is enough to make it a challenge even to find many more than 100 or so cognates that survive (and can still be straightforwardly elicited) in all of the dialectal variation within each branch. Details on this, and on other criteria to ensure that the list makes for as balanced as possible a sample of the phonetics of each language family, are given in [1].

5. TRANSCRIPTION POLICIES AND NORMALISATION TO CLTS

For any database of phonetic transcriptions, key concerns are consistency and standardisation. All *Sound Comparisons* transcriptions follow IPA usage, in Unicode. Even so, in practice those still leave much scope for inconsistency. The need for data to be consistent, standardised and unambiguous is particularly acute for the growing trend towards analysing phonetic transcription data by computational methods.

Sound Comparisons seeks to limit transcription inconsistencies through specific conventions and guidelines for its transcribers, and by ensuring that wherever possible, all transcriptions within each family, or at least major sub-branch, are by the same phonetician. Transcriptions are generally narrowest, though, for English, Mapudungun, Romance and Balto-Slavic; and broader for Germanic, the Andean families, and particularly Vanuatu.

All *Sound Comparisons* transcriptions have also been run through computer-assisted correction and normalisation. The Cross-Linguistic Transcription System (CLTS, see [2]) provides software that can identify many forms of error or departures from IPA and Unicode usage, and can correct transcriptions to a proposed 'normalised' IPA. CLTS tools identify and correct Unicode "confusables" (see [3]), e.g. [ə] U+0259 "Latin small letter schwa" *vs*. [ə] U+01DD "Latin small letter turned e", and common transcription 'shorthand' characters, e.g. the ASCII colon [:] instead of the correct Unicode length marker [ː] U+02D0. CLTS also normalises the order of diacritics, e.g. [kʲʰ] rather than [kʰʲ], and their best placement relative to the base symbol (e.g. a devoiced ring either above or below it). In return, the narrow phonetic transcriptions of *Sound Comparisons* served as an extensive test that contributed to the final stage of development, refinement and extension of the CLTS software itself.

6. WEBSITE RESEARCH FUNCTIONALITY: SEARCH, FILTER, DOWNLOAD, CITE

The two axes of the database, languages and words, also structure the explorer website. The main central panel is flanked by a language selector panel on the left, and a word selector panel to the right. The data thus selected appear in the main central panel, in map or table views, showing: the pronunciations of a single selected word in all languages; of all words in a single language; or of any selected combination of words and languages. The corresponding sound files are played and compared instantaneously by touching, clicking or just hovering the cursor over any transcription or play icon.

Among many search and filter options, most useful for phonetic research are two entry boxes in the word selector column: to filter/search by either the orthography or the phonetic transcriptions of any language variety in that study. So a user can set the spellings filter language to English, and type f to return all words that contain the <f> grapheme in their English spelling; or set the transcriptions filter language to Doric Scots, and type f to return all words that contain IPA [f] in their transcriptions in Doric Scots. All cognate reflexes in other languages can then be shown by just clicking an icon to reset the main display table to this newly filtered set of words (in the multi-language table views).

Both filter boxes allow full 'regular expressions' syntax. So typing ^f or f$, for instance, returns all words that begin or end with f, respectively. The transcription filter box has pop-up IPA charts so that the user can enter any IPA character or diacritic directly, without needing complex key combinations. It also recognises phonological shorthand characters in upper case: i.e. V for any vowel, N for any nasal, and so on. Entering VːN$, for example, returns all words that end in a long vowel followed by a nasal. It is possible to filter for any Unicode character, even if only a modifier or diacritic, to find all instances, irrespective of the main symbol it appears with. So typing only ʷ returns all words with any labialised segment. With the search language set to an earlier, ancestral language, the user can filter to a given sound in that language, and then immediately create a side-by-side table to compare all of that (proto-)sound's reflexes across all descendant varieties, e.g. all modern Romance reflexes of Classical Latin /kʷ/ or written <qu>, or all modern reflexes of Early Modern English <r>.

For any combination of languages and words customised using these search and filter operations, clicking on the link icon gives a durable shortcut URL so that the user can also cite that specific combination of data. Clicking on the download icons

exports to the user the phonetic transcriptions and the sound files for that precise selection of languages and words. Exports are in Unicode plain csv and tsv formats, and comply with the proposed Cross-Linguistic Data Format (CLDF) standards (see [4]), to facilitate integration with the suite of databases that already follow this format (clld.org/datasets.html). This progressive integration aims also to ensure the sustainability and long-term data curation that CLDF and CLLD are specifically devised to provide.

It is also possible to cite any studies and views with custom URLs after the soundcomparisons.com/ root (replaced by ../ in the links below), as follows.

- Any specific study: e.g. ../Germanic.
- Any specific view type and word: e.g. ../Germanic/map/night
- Any current user interface language for the website (English, Bislama, Croatian, German, Italian, Polish, Portuguese, Russian, Spanish), using its two-letter ISO 639-1 code: e.g. ../bi/Vanuatu for the Bislama version.
- All data for any one language variety, using its three-letter ISO 639-3 code, or its aaaa0000 format GlottoCode: e.g. either ../cap or ../chip1262 for the Chipaya language.
- Often a single ISO or GlottoCode has several sub-varieties in *Sound Comparisons*, so those codes return a table of all of those sub-varieties: e.g. ../vls shows the three varieties that fall under ISO code vls for Flemish.
- Any individual variety is specified by its *Sound Comparisons* URL-name: e.g. ../Gmc_W_Dut_ZandF_FlW_FrenchFlemish_Dl for the 'Flemish' specific to northern France.

7. RANGE OF RESEARCH APPLICATIONS

Some of the earliest publications founded on the data now built into *Sound Comparisons* focused on English dialectology (see [5], [6]). Those were soon extended to Germanic historical linguistics, to the general methodological challenge of quantifying linguistic distance, and to whether such measures can validly contribute to diachronic study, as in [7].

On a vexed question of language classification, meanwhile, [8] addresses whether Huilliche qualifies as either a fully-fledged language and 'sister' to Mapudungun, or just a slightly divergent dialect within it. *Sound Comparisons* data were employed as evidence to clarify the nature and degree of phonetic divergence that underlies the classificatory confusion. Mapudungun is also characterised by a "typologically famous series of interdental-alveolar oppositions", i.e. not just /θ/ *vs*. /s/, but also /t̪/ *vs*. /t/, /n̪/ *vs*. /n/ and /l̪/ *vs*. /l/. Despite claims of its demise, in [9] this opposition is shown to survive in various of the 37 regiolects of Mapudungun covered by *Sound Comparisons*, as at ../sl/2G (where /sl/ = a short link). Other data such as at ../sl/n5, meanwhile, are employed in [8] to show that one Huilliche variety may even have developed a "bisyllabic consonant cluster with phonemic status /l̪d̪/".

Sound Comparisons is also particularly suited to research on dialect continua, including where migration and contact bring further complexities. As one illustration, the western Erzgebirge, the 'Ore Mountains' that form a natural frontier between Germany and the Czech Republic, were from the 12th century onwards settled primarily by speakers of High Franconian and Thuringian dialects, who thus also brought their speech into contact with new, more easterly neighbours. Even a small customised sample at ..sl/Mm can reveal this mix of origins and contact. Western Erzgebirgisch deletes word-final *n (as in 'stone' in Table 1) and intervocalic *g (as in 'nail'), both traits probably inherited from East Franconian. But it also shows the typically (East) Thuringian hardening of word-initial *j- (as in 'year'), while its vowel system has been re-shaped by retracting *a (as in 'name') and by fronting back rounded vowels (as in 'dog'), traits typical of Upper Saxon, the main contact variety since the migrations.

Table 1: Western Erzgebirgisch compared to its source and contact dialects — data at ..sl/Mm.

dialect group	East Franconian	Thuringian	Upper Saxon	West Erzgebirge
locality	Altersbach	Altenburg	Leipzig	Aue
'stone'	ʃtaɪ	ʃteːn	ʃtɛɪn	ʃtæː
'nail'	nɛːl	nɔ̰ːʁəl	nʌːɣəl	nʌːəl
'year'	juwə̆	ɡɔˑʌ	jɔˤː	ɡɔ̰ˤː
'name'	nuːmə	nɤːmə	nʌːmə	nʌːmə
'dog'	hɔʏnd̰	hʌntʰ	hønt	hønt

Other language families and regions covered within *Sound Comparisons* offer many further data-sets open to researchers to exploit: on, for instance, debated aspects of English dialectology such as 'pre-glottalisation' in Tyneside English; on geminate reflexes across Italian dialects; on prenasalisation phenomena in many near-undocumented Oceanic languages of Vanuatu; and on the striking consonant clusters in the little-known Chipaya language of highland Bolivia, at ../cap.

In sum, the scale, diversity and precision of the *Sound Comparisons* database, coupled with its powerful research functionality and full open access online, make for a rich resource. Researchers in phonetics are invited to explore and make use of it.

AUTHORS, AFFILIATIONS AND EMAILS

Paul Heggarty[1], Aviva Shimelman[2], Giovanni Abete[3], Cormac Anderson[1], Scott Sadowsky[4,1], Ludger Paschen[5], Warren Maguire[6], Lechoslaw Jocz[7], María José Aninao A.[8], Laura Wägerle[2], Darja Appelganz[1], Ariel Pheula do Couto e Silva[9], Lewis C. Lawyer[2], Ana Suelly Arruda Câmara Cabral[9], Mary Walworth[1], Jan Michalsky[10], Ezequiel Koile[1], Jakob Runge[2] & Hans-Jörg Bibiko[1]

[1] Dept of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany
[2] Independent scholar. (*Sound Comparisons* work performed during employment at [1] and/or at the Max Planck Institute for Evolutionary Anthropology, Leipzig.)
[3] Dipartimento di Studi Umanistici, University of Naples Federico II
[4] Dpto. de Ciencias del Lenguaje, Pontificia Universidad Católica de Chile, Santiago
[5] Leibniz-Zentrum Allgemeine Sprachwissen-schaft, Berlin
[6] Linguistics and English Language, University of Edinburgh
[7] Akademia im. Jakuba z Paradyża, Gorzów Wlkp., Poland
[8] Pontificia Universidad Católica del Perú, Lima
[9] Laboratório de Línguas e Literatura Indígenas, University of Brasilia
[10] Technology Management, FAU Erlangen-Nuremberg

Paul.Heggarty@gmail.com, nomdecrayon@gmail.com, giovanni.abete@unina.it, cormacanderson@gmail.com, ssadowsky@gmail.com, ludger.paschen@rub.de, lechoslaw.jocz@gmail.com, mj.aninao.a@gmail.com, laura.waegerle@posteo.de, darjaappelganz@googlemail.com, ariel.bsaz@gmail.com, lclawyer@ucdavis.edu, asacczoe@gmail.com, jan.michalsky@fau.de, ezequielk@gmail.com, runjak@gmail.com, mail@bibiko.de

ACKNOWLEDGEMENTS

We thank the many hundreds of native speakers who generously gave of their time and allowed us to record their voices, as well as all other contributors to the *Sound Comparisons* project.
This project has relied on extensive funding from the British Academy, Leverhulme Trust and Max Planck Society. Full details are given at: soundcomparisons.com/#/about/Funding

REFERENCES

[1] Heggarty et al. in prep. Sound Comparisons: A new resource for exploring phonetic diversity across language families of the world.
[2] Anderson, C. et al. 2019. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting* 4.
[3] Moran, S., & Cysouw, M. eds. 2018. *The Unicode Cookbook for Linguists.* Berlin: Language Science Press. http://doi.org/10.5281/zenodo.1296780
[4] Forkel, R. et al. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5: p.180205. http://doi.org/10.1038/sdata.2018.205
[5] McMahon, A.M.S. et al. 2007. The sound patterns of Englishes: representing phonetic similarity. *English Language and Linguistics* 11 (01): p.113-42. http://doi.org/10.1017/S1360674306002139
[6] Maguire, W. et al. 2010. The past, present and future of English dialects: Quantifying convergence, divergence and dynamic equilibrium. *Language Variation and Change* 22 (1): p.69-104. http://doi.org/10.1017/S0954394510000013
[7] Heggarty, P., Maguire, W., & McMahon, A.M.S. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences* (365): p.3829-43. http://doi.org/10.1098/rstb.2010.0099
[8] Sadowsky, S. et al. 2015. Huilliche: ¿geolecto del mapudungun o lengua propia? Una mirada desde la fonética y la fonología de las consonantes. In: A. Fernández Garay & M. A. Regúnaga (eds) *Lingüística indígena sudamericana: aspectos descriptivos, comparativos y areales.* Buenos Aires: Editorial de la Facultad de Filosofía y Letras, Universidad de Buenos Aires.
[9] Sadowsky, S. et al. 2013. Mapudungun. *Journal of the International Phonetic Association* 43 (01): p.87-96. http://doi.org/10.1017/S0025100312000369