Dmitry Nikolaev*

The Database of Eurasian Phonological Inventories: a research tool for distributional phonological typology

https://doi.org/10.1515/lingvan-2017-0050
Received November 10, 2017; accepted February 12, 2018

Abstract: The paper presents an overview of *The Database of Eurasian Phonological Inventories*—a new information resource and analytical tool for research in the field of distributional phonological typology, theoretical phonology, and areal linguistics.

Keywords: Eurasia; database; areal linguistics; segment inventory; phonological typology.

1 Introduction

The aim of this paper is to present to the research community *The Database of Eurasian Phonological Inventories* (EURPhon)[1]—an information resource dedicated to the phonological inventories of the languages of Eurasia. The main goal of the project is to provide access to data about the phonological inventories of Eurasian languages (enriched where possible by the additional phonotactic information), both in the form of a user-friendly online database with novel query mechanisms, on the one hand, and in a structured format amenable to statistical processing, on the other.

The structure of this paper is as follows: § 2 surveys available phonological databases and explains the rationale for the project; § 3 describes the data extraction methodology followed in EURPhon; § 4 outlines the structure of database entries; § 5 describes EURPhon's online interface, the structure of the downloadable data-file, and the freely available software powering the database, which can be used as a stand-alone research tool; § 6 presents a case-study of the accumulation of place-manner distinctions in the consonant inventories of the languages of Eurasia; § 7 concludes.

2 Existing phonological databases and the scope of the project

The first decade and a half of the 21st century have been a period of remarkable activity in the field of phonological databases, which has long been dominated by the seminal work of Ian Maddieson (Maddieson 1984; Maddieson and Precoda 1992). New databases focused mostly or exclusively on segmental inventories include Merritt Ruhlen's database published as Supplementary Materials to (Creanza et al. 2015), Jeff Mielke's P-base (http://pbase.phon.chass.ncsu.edu/query), Lyon-Albuquerque Phonological Systems Database (LAPSyD, http://www.lapsyd.ddl.ish-lyon.cnrs.fr/), PHOIBLE (Moran et al. 2014), and SAPhon (Michael et al. 2015). Information on segmental inventories is also included in the World Phonotactics Database (Donohue et al. 2013; although it does not provide access to the raw data and only permits a wide range of predefined types of queries).[2]

[1] http://eurasianphonology.info/

[2] A huge amout of work has also been done on creating databases of cognate sets, which can be used to automatically extract information about phonemic inventories and sound-change processes. Major projects include Benchmark Database for Phonetic Alignments (http://alignments.lingpy.org/), A Comparative Study of Andean Languages (http://quechua.org.uk/Eng/Cpv/), and The Global Lexicostatistical Database (http://starling.rinet.ru/new100/main.htm). The UniDia project (http://www.diadm.ish-lyon.cnrs.fr/unidia/) is dedicated to cataloguing recorded and reconstructed sound-change processes.

***Corresponding author: Dmitry Nikolaev,** Dynamics of Language Lab, The Hebrew University of Jerusalem, Jerusalem, Israel, E-mail: dmitry.nikolaev@mail.huji.ac.il. http://orcid.org/0000-0002-3034-9794

Altogether, these resources provide impressive coverage of the phonologies of the world's languages, which has made it possible to conduct statistical investigations into the extralinguistic correlates of different phonological configurations (see Ladd et al. 2015 for an overview). What is lacking in most of these projects, however, with the notable exception of SAPhon, is the complete coverage, or at least a highly dense sample of a given macro-area, which would make it possible to successfully account for different kinds of areal and genetic effects in the investigation of distributions of phonological features. The importance of such analyses has repeatedly been stressed by scholars working in the field of distributional typology (Bickel 2015), and they may help to resolve the problematic issue of linguistic areas (Daumé III 2009; Campbell 2017). Whereas the SAPhon project provides such densely-sampled data for South American languages, the EURPhon project aims to cover Eurasia.

A recent estimate of the number of languages currently spoken in Eurasia puts the number at close to 1500 (Hammarström and Donohue 2014); however, adequate phonological descriptions seem to exist for only about 600 of them. At the moment, data on 388 languages are available in the online version of the database with new data being added continuously.

3 Data collection and extraction

In the framework of the project, *Eurasia* is understood as including Atlantic islands up to Iceland and those along the eastern and southeastern coast of the main landmass, but excludes the Indonesian archipelago and the Philippines.

The data for the project are extracted from descriptions of individual language varieties contained in grammars, phonological analyses, and descriptive sketches published independently or as parts of overview works on different language families or regions. Data from existing databases are not reused, and no first-hand analyses of raw data (such as word-lists and text collections) are undertaken.

There are two type of data-points in the database: *languages* and *dialects*. At the moment the data on dialects is available only for download and is not included in the maps, reports, and in the search results. When choosing which varieties to label as *languages* as opposed to *dialects*, the main criterion is that of mutual intelligibility. Thus, when confronted with a group of varieties, which are described as mutually intelligible, one of them (usually the best described one) is chosen as a *language*-entry and others are marked for inclusion as *dialects*. In some cases, such data are not available, and the decision is made based on labels assigned by the authors of respective descriptions.

When presented with conflicting descriptions of the same varieties, the general approach is to use the one most rich in phonetic detail and/or using the most IPA-compliant notation. In most cases, careful descriptions are in good agreement with each other, in which case the most detailed or the most recent one is used. Sometimes, however, there are several heavily phonologised or impressionistic descriptions of the same variety (or descriptions using uninformative traditional notation such as is common in Turkic, Slavic, or Uralic studies). In these cases, phonetic details from several descriptions are used where possible in order to provide an IPA interpretation of the system.

All the segmental data are entered in the database in IPA notation, which insures comparability and makes it possible to construct feature-based queries (cf. § 5 below). Phonological analyses contained in the sources are mostly followed and the notation is preserved as far as possible. Deviations are systematically warranted in the following cases:

- When the major allophone of a phoneme is effectively hidden by a simplified notation, e.g., when the opposition of plain and aspirated voiceless stops is represented as an opposition of voiced vs. voiceless stops or when /a/ is systematically used to represent /ɑ/.
- When a complex segment well established for the languages of a region is represented as a sequence of segments for the sake of inventory simplicity, e.g., when a pre-nasalised segment is analysed as a sequence of 'a nasal archiphoneme' or 'a nasalisation phoneme' followed by a consonant.

The overall principle is that the notation used should straightforwardly represent the phonetic properties of major allophones of proposed phonemes. In order to avoid undue dependence on phonemic analyses, sometimes rather radical, proposed by different scholars, a unified approach was adopted where an allophone of a phoneme found in a word-initial position is taken as its main representative (except for unstressed vowels in this position; the basic context for vowels is in a stressed syllable, preferably word initially or after a neutral consonant such as /p/). This allowes for a unified treatment of most described inventories as the information on consonant allophones in this position is usually provided. A more cautious approach would entail analysis of all positions in which a given phoneme can be found, which is in most cases impossible due to the scarcity of data.

Up to this point, two cases were found where the deviation from IPA notation seemed warranted:

1. The so-called hissing-hushing fricatives of Northwest Caucasian languages (denoted by /ŝ ẑ/ following Catford 1977; Ladefoged and Maddieson 1996).
2. The so-called apical vowels found in Mandarin Chinese and some other East Asian languages.

In both cases, there is no agreed-upon IPA description of these sounds as their exact articulation is still debated, but there is a stable descriptive practice in the respective descriptive traditions, which does not lead to notational conflicts.

4 Structure of the database entry

Database entries obligatorily include (i) the name of the language, (ii) geographical coordinates (which should ideally represent the center of the geographical distribution of the described variety), (iii) consonant and vowel inventories (including diphthongs and triphthongs), (iv) phylum membership, (v) the bibliographical description of the source of the data, and (vi) the name and email address of the person who submitted the data.[3]

The following information is provided when it is relevant and available: the ISO code, genus membership, tone inventory, attested syllable types, initial clusters, single consonants and clusters found in the syllable-final position.

Entries also include a ‘Comments’ section pointing out problematic aspects of descriptions and gaps in them, as well as documenting cases where data have been re-analysed. The presence of phonemes whose distribution is restricted to recent loans is also usually recorded in this section.

5 Data presentation and query mechanisms

In this section, we first describe the online interface to the web-based version of EURPhon (http://eurasianphonology.info/), then describe the structure of the downloadable data file, and finally briefly discuss the programming interface of the software powering EURPhon, which can be used as a standalone research tool.

5.1 The online interface

The online interface of EURPhon provides three views of the database: a mapview, a listview, and a segment view.

[3] The lion’s share of the entries were added by Dmitry Nikolaev, who is also responsible for checking all the data.

The **mapview** shows all the languages on the map, with colours of the points corresponding to phyla. Entries for individual languages can be accessed by clicking on the markers. The screenshot of the mapview is shown in Figure 1, and the inventory of Sindhi in Figure 3 in the Supplementary Materials. The **listview** shows the languages organised according to their genealogical affiliation and in the alphabetical order. A two-tier description consisting of 'family' (~Indo-European) and 'group' (~Slavic, Germanic) is used. The **segment view** presents all the segments that can be found in the languages in the database. The distribution of each segment can be accessed by clicking on it. A part of the segment view is presented in Figure 2 in the Supplementary Materials.

The **family/group reports** section provides information about particular phyla and genera, i.e., the geographical distribution of languages from the group included in the database, common phonemes, histograms of total segment counts and consonant and vowel counts. As an example, the report on the Nakh-Daghestanian family is presented in Figure 5 in the Supplementary Materials.

The heart of the database is the **search** section. At the moment, the database accepts three types of queries:

1. Exact phoneme queries
2. Fuzzy phoneme queries
3. Feature queries

Exact phoneme search: Returns a list of languages that have a particular segment encoded by IPA, e.g. /p/ or /ə̃ː/. It is also possible to search for inventories with gaps. For example, the query 'p' will return the list of languages that have /p/, while the query '-p' will return the list of languages that do not have /p/. It is also possible to make composite queries combining positive and negative elements separated by commas: 'a, -b, c^{h}, -d^{j}'.

Fuzzy phoneme search: Returns the variants of the input phoneme containing additional privative IPA features and their distributions. For example, the query 't' will return the distribution of the basic variant /t/, as well as those of /t^{h}/, /t^{hj}/, /t^{hw}/, /^{h}tː/, etc. It must be noted that fuzzy search operates on the level of IPA features, so different input sequences may be parsed in the same way. For instance, voiceless segments in several languages are described as 'unvoiced' /d̥ b̥ g̥/ in order to better capture their phonetic properties. This feature has not been as yet implemented in the database, and fuzzy queries for the variants of /t/ return /d̥/ and vice versa.

Feature search: Makes it possible to query for inventories having or lacking segments characterised by particular bundles of IPA features. For example, it is possible to query for languages that do not have laterals ('-lateral'), have lateral fricatives ('lateral fricative'), or have lateral fricatives, but lack lateral affricates ('lateral fricative, -lateral affricate'). The distribution returned by the search system for the query 'pharyngealised plosive' is shown in Figure 4 in the Supplementary Materials.

5.2 Downloadable data and software

The search mechanisms described in the previous section enable complex investigations into the make-up of Eurasian phonological inventories; however, they are still rather restrictive, and it is not possible to directly incorporate them in a statistical-analysis pipeline.

In order to facilitate statistical research on Eurasian phonologies and to make it possible to construct queries of unrestricted complexity, the full dataset of EURPhon is available for download as a JSON file.[4]

Each language in the database is indexed by a unique ID. Language entries themselves are dictionaries with a fixed set of keys:

[4] JSON (http://www.json.org) is an hierarchical human-readable data format allowing for string-indexed array-like entries of variable length.

1. ‘code’ (ISO code)
2. ‘type’ (language vs. dialect)
3. ‘name’
4. ‘coords’ (a 2-tuple consising of latitude and longitude)
5. ‘gen’ (a 2-tuple consisting of phylum and genus)
6. ‘inv’ (an array of consonants and vowels)
7. ‘cons’ (an array of consonants)
8. ‘vows’ (an array of vowels)
9. ‘tones’ (an array of tones described using tone numerals with some additional notation for breathy, creaky, and checked tones: 44, 44ɦ [breathy], 44ʕ [creaky], 4ʔ4 [interrupted], 44ʔ [checked])
10. ‘syllab’ (a string describing attested syllable structures)
11. ‘cluster’ (a string describing attested initial clusters)
12. ‘finals’ (a string describing attested finals and final clusters)
13. ‘source’
14. ‘comment’
15. ‘contr’ (name and email address of the person who submitted the data)

The JSON data format is supported by all major programming languages including Python, the language, in which the EURPhon engine is written.[5]

In order to enable feature-based queries, a software library `IPAParser` was developed. The `IPAParser` library exports the `parsePhon` function, which takes as input an IPA-encoded phoneme (such as ‘p’) and outputs its IPA-feature make-up ({‘voiceless’, ‘plosive’, ‘bilabial’}). By iterating over inventories and segments in them it is possible to construct feature queries of arbitrary complexity.

Finally, the search engine powering the database is itself made available as the library `PhonoSearchLib`, which exports the `LangSearchEngine` class with methods `IPA_exact_query`, `IPA_query_multiple`, `IPA_query` (corresponding to the fuzzy phoneme search), and `features_query`.[6] A LangSearchEngine is initialised with a JSON data-file, which can be downloaded as a part of the repository or from the EURPhon web-site. Using these libraries, it is possible to incorporate the EURPhon data and search capabilities into any kind of statistical analysis pipeline.

6 A case study: accumulation of place-manner combinations in the languages of Eurasia

In order to give a simplified example of such a pipeline, we present a small case-study dealing with the accumulation of place-manner distinctions for consonants in the languages of Eurasia. This case-study builds on work by Lindblom and Maddieson (1988), who found, based on a typological sample, that in the process of consonant-inventory growth languages tend to exhaust basic place-manner combinations by means of additional articulations before acquiring new place-manner distinctions.

In order to check this result, we counted the number of different place-manner combinations in consonants for all languages in the database. A scatterplot of the dependence of the number of different place-manner combinations in a given language on the total number of consonants in it with LOESS smoothing is shown in Figure 1. It is evident that there is a clear linear relationship between the number of consonants in an inventory and the number of different place-manner distinctions, although there is much variation.[7]

[5] In the future, it is planned to migrate the database to the SQL format and change the data dump accordingly.

[6] The source code for the database engine with all component libraries is maintained in a GitHub repository: https://github.com/macleginn/eurasian-phonologies

[7] This dependence is significant with $p < 0.001$ in a mixed model with random intercepts for phyla.

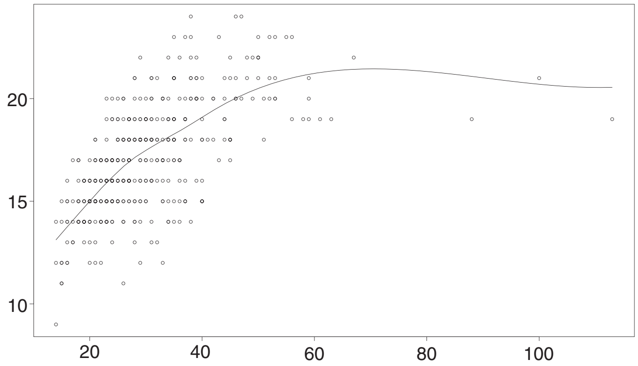

Figure 1: Dependence of the number of different place-manner distinctions on the number of consonants in the languages of Eurasia

Moreover, we see an upper-bound effect: languages with super-saturated consonant inventories (60+ segments) do not acquire new place-manner distinctions compared to languages with c. 40 different consonants. Python and R scripts used to prepare the data, fit the mixed linear model, and produce the plot in Figure 1 are presented in the Supplementary Materials.

7 Conclusion

EURPhon is still undergoing active development. New features, such as queries involving counts of segments of particular types, are slated for inclusion; new data are being added continuously, and old entries are being checked and updated when new descriptions become available. However, EURPhon is already a powerful research tool, which can be used for distributional phonological typology, areal and contact studies, and correlational investigations.

The online interface of the database provides several advanced and novel query mechanisms, but the real power of EURPhon lies in the combination of open data and rich IPA-parsing and search libraries, which are distributed together with the data and make it possible to easily conduct feature-based queries of arbitrary complexity.[8]

Acknowledgment: The author would like to thank Eitan Grossman and two anonymous reviewers for their careful reading of the manuscript and valuable comments and suggestions. The writing of this article was supported by a post-doctoral grant from the Dynamics of Language Lab of the Hebrew University of Jerusalem.

References

Bickel, Balthasar. 2015. Distributional typology. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 901–923. Oxford: Oxford University Press.

Campbell, Lyle. 2017. Why is it so hard to define a linguistic area? In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 19–39. Cambridge: Cambridge University Press.

Catford, John C. 1977. Mountain of tongues: The languages of the Caucasus. *Annual Review of Anthropology* 6(1). 283–314.

Creanza, Nicole, Merritt Ruhlen, Trevor J. Pemberton, Noah A. Rosenberg, Marcus W. Feldman & Sohini Ramachandran. 2015. A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences* 112(5). 1265–1272.

[8] The database release also includes the `IPATabulator` library, which automatically constructs HTML-formatted tables from arrays of IPA-encoded phonemes.

Daumé III, Hal. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, 593–601. Association for Computational Linguistics.

Donohue, Mark, Rebecca Hetherington, James McElvenny & Virginia Dawson (eds.). 2013. *World phonotactics database*. Canberra: Department of Linguistics, The Australian National University. http://phonotactics.anu.edu.au.

Hammarström, Harald & Mark Donohue. 2014. Some principles on the use of macro-areas in typological comparison. *Language Dynamics and Change* 4(1). 167–187.

Ladd, D. Robert, Seán G. Roberts & Dan Dediu. 2015. Correlational studies in typological and historical linguistics. *Annual Review of Linguistics* 1(1). 221–241.

Ladefoged, Peter & Ian Maddieson. 1996. *The sounds of the world's languages*. Oxford, OX, UK and Cambridge, Mass., USA: Blackwell Publishers.

Lindblom, Björn & Ian Maddieson. 1988. Phonetic universals in consonant systems. In Victoria Fromkin, Larry M. Hyman & Charles N. Li (eds.), *Language, speech, and mind: studies in honour of Victoria A. Fromkin*, 62–78. London and New York: Routledge.

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

Maddieson, Ian & Kristin Precoda. 1992. *UPSID and PHONEME (version 1.1)*. Los Angeles: University of California at Los Angeles.

Michael, Lev, Tammy Stark, Emily Clem & Will Chang (eds.). 2015. *South American phonological inventory database v1.1.4*. Berkeley: University of California. http://linguistics.berkeley.edu/~saphon/.

Moran, Steven, Daniel McCloy & Richard Wright (eds.). 2014. *Phoible online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://phoible.org/.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/lingvan-2017-0050).