Defining and Counting Phonological Classes in Cross-linguistic Segment Databases

Dan Dediu, Scott R. Moisik
Max Planck Institute for Psycholinguistics
Wuldtlaan 1, Nijmegen, The Netherlands
Dan.Dediu@mpi.nl, Scott.Moisik@mpi.nl

Abstract

Recently, there has been an explosion in the availability of large, good-quality cross-linguistic databases such as WALS (Dryer & Haspelmath, 2013), Glottolog (Hammarström et al., 2015) and Phoible (Moran & McCloy, 2014). Databases such as Phoible contain the actual segments used by various languages as they are given in the primary language descriptions. However, this segment-level representation cannot be used directly for analyses that require generalizations over classes of segments that share theoretically interesting features. Here we present a method and the associated R (R Core Team, 2014) code that allows the flexible definition of such meaningful classes and that can identify the sets of segments falling into such a class for any language inventory. The method and its results are important for those interested in exploring cross-linguistic patterns of phonetic and phonological diversity and their relationship to extra-linguistic factors and processes such as climate, economics, history or human genetics.

Keywords: phonetics, phonology, segment classes

1. Introduction

The patterns of phonetic and phonological diversity are very important to understanding the wider linguistic diversity, the processes shaping it and its evolution. For several decades now, databases such as the *UCLA Phonological Segment Inventory Database* (UPSID) (Maddieson, 1984) have played a major role in allowing (semi-)quantitative analyses of the patterns of phonetic and phonological variation around the world. Unfortunately, until relatively recently, such databases used to be relatively small (for example, UPSID has 919 different segments in 451 languages; `http://web.phonetik.uni-frankfurt.de/upsid_info.html`) and offer a rather sparse geographical and genealogical coverage. The recent advent of the *World Atlas for Language Structures* (WALS) (Haspelmath et al., 2005) and its online and continuously updated version (Dryer and Haspelmath, 2013) available at `http://wals.info`, offer better coverage (currently 2,679 languages, many with lots of missing data) but at the cost of providing only a small number of variables (currently only 20 are listed under "Phonology"; `http://wals.info/feature`) that are discretized into a limited number of categories (e.g., feature 1A "Consonant inventories" can only have five values: "small", "moderately small", "average", "moerately large", and "large"), a representation that is hard to accomodate in quantitative analyses and whose use has justifiably generated intense controversy. Fortunately, the last several years have seen the release of phonetic/phonological databases that are much richer and have better coverage, such as *Phoible*[1] (Moran et al., 2014) and the dataset accompanying a recent comparison of phonological and genetics patterns of diversity (Creanza et al., 2015) curated by Merritt Ruhlen[2]. However, such segment-level databases have a major drawback intrinsic to their design in that they cannot be directly used for analyses that require generalizations over *classes* of segments that share theoretically interesting features, such as "front rounded vowels", "retroflex stops" or "clicks".

In this paper we[3] introduce a method that builds on the notion that a *system of atomic features* can be used to describe each and every possible segment in such databases, allowing the flexible definition of classes of segments[4]. These classes can then be applied to any given language in the database: the sets of segments that fall into these classes can be identified and counted (and, further on, discretized into a small number of bins or even binarized as presence/absence), resulting in a higher-level representation of the phonetic and phonological systems of the language. These more general classes (such as "vowel", "consonant" or "click") allow then the quantitative exploration of more abstract patterns of cross-linguistic diversity and their relationship to extra-linguistic factors, among other applica-

[1] Freely available at `http://phoible.org`.

[2] Freely available at `http://www.pnas.org/content/suppl/2015/01/15/1424033112.DCSupplemental/pnas.1424033112.sd01.txt` and `http://www.pnas.org/content/suppl/2015/01/15/1424033112.DCSupplemental/pnas.1424033112.sd02.txt`.

[3] Author contributions: DD and SRM desgined the research; SRM defined the *Fonetikode* feature system and applied it to the *Phoible* and *Ruhlen* databases; SRM defined the segment classes and their specification in both feature systems; DD wrote the `R` script implementing the classes and generated the UULIDs; DD and SRM checked the results; DD wrote the paper.

[4] In this paper we do not commit ourselves to any particular such system or proposed justification for their existence, treating such systems as user-controlled parameters.

tions. As an example, we can investigate the distribution and historical dynamics of "retroflex stops" (consonants such as [t̠] and [ɖ̠]) in the world's languages using statistical and phylogenetic methods.

The `R` (R Core Team, 2014) code that implements this method and the input (the raw data as provided by two segment-level databases) and output files (containing, for each language in the databases, the composition and count of number of segments in each class) are freely available under liberal licenses[5] on the GitHub repository `https://github.com/ddediu/phon-class-counts`, where details about the file formats and the implementation are in the `README.md` file. We hope that these data, the script, and the results will be used in large-scale, cross-disciplinary, data-driven statistical explorations of linguistic diversity and its causes.

2. Data and Methods

This section describes the two databases and the two feature systems we used, as well as the procedures for defining classes of segments and identifying and counting them for a given language inventory.

2.1. The Databases

Phoible (Moran et al., 2014) is a very large, freely available, and continuously evolving database that contains (in the 2014 edition used for this paper) 2,155 phonological inventories for 1,672 distinct languages (some of these languages can have more than one description available) covering 2,160 segment types; in the following we will denote this database as *Phoible* or *P*.

Recently, Creanza et al. (2015) have released as Supplementary Materials Online (Datasets S1 and S2) a database collected and curated by Merritt Ruhlen containing the presence/absence of 728 segment types[6] in 2,082 languages (please note that the transcription system used by this database is quite different from the transcription used by Phoible); we will denote this database as *Ruhlen* or *R*.

These two databases are complementary, in the sense that they describe overlapping but different sets of languages, and the actual inventories of the same language sometimes show differences.

[5]The `R` script is released under GPL version 3 (`http://www.gnu.org/licenses/gpl.html`). The input files created by us are released under Creative Commons Attribution 4.0 International (CC BY 4.0; `http://creativecommons.org/licenses/by/4.0/`), as are all output files. The input files that do not belong to us are governed by their respective licenses: Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0; `http://creativecommons.org/licenses/by-sa/3.0/`) for the *Phoible* data and by the PNAS terms (`http://www.pnas.org/site/misc/terms.xhtml`) for the *Ruhlen* database.

[6]After pre-processing (such as removing symbols that seem to stand for more abstract categories, such as "c" and "v" and their variants, or "vowelharmony"), there are 644 potential segment types.

2.2. The Feature Systems

The *Phoible* database also provides a description of its segments in terms of four possible values, "+", "–", "0" and "" (absent), on 37 features such as "tone", "syllabic", "short" and "continuant" (for details see `http://phoible.org`). This feature system is described as being "[...] loosely based on (Hayes, 2009) with some additions drawn from (Moisik and Esling, 2011)", and we will denote it in the following as *Phoible* or *P* (with explicit disambiguation from the segment database only when necessary).

We also constructed a feature system of a more phonetic nature inspired from the International Phonetic Alphabet or IPA (International Phonetic Association, 2005) containing 13 features, each with its own set of possible values; we will denote this system as *Fonetikode* or *F*. These features (**bold**) and their possible values (*italic*) are: **GC** (general class): *c* (consonant), *v* (vowel), *ss* (suprasegmental/tone); **VV** (vowel vertical or tongue height): *c* (close), *nc* (near close), *cm* (close mid), *m* (mid), *om* (open mid), *no* (near open), *o* (open); **VH** (vowel horizontal or tongue anteriority): *f* (front), *nf* (near front), *c* (central), *nb* (near back), *b* (back); **VM** (vowel modifiers): *none*, *n* (+nasal), *r* (+round), *nr* (+nasal to +round), *ur* (–round to +round), *ru* (+round to –round), *un* (–nasal to +nasal), *nu* (+nasal to –nasal), *nur* (nasalized –round to +round), *nru* (nasalized +round to –round); **CP** (consonantal place of articulation): *b* (bilabial), *bld* (bilabio-labiodental), *ba* (bilabio-alveolar), *bpa* (bilabio-postalveolar), *ld* (labiodental), *lv* (labiovelar), *lu* (labiouvular), *ll* (linguolabial), *i* (interdental), *d* (dental), *a* (alveolar), *pa* (postalveolar), *r* (retroflex), *dv* (dentivelar), *av* (alveolovelar), *ap* (alveolopalatal), *p* (palatal), *pav* (postalveolar-velar), *v* (velar), *u* (uvular), *ph* (pharyngeal), *e* (epiglottal), *g* (glottal); **CM** (consonantal manner of articulation): *s* (stop), *f* (fricative), *af* (affricate), *n* (nasal), *a* (approximant), *t* (tap/flap), *tr* (trill), *clk* (click); **CS** (consonantal sequencer): *s* (simplex), *c* (complex), *prn* (pre- nasalized), *pon* (post-nasalized), *prs* (pre-stopped), *pos* (post-stopped), *prg* (pre-glottalized), *pog* (post-glottalized), *pra* (pre-aspirated), *poa* (post-aspirated), *prag* (pre-aspirated & glotatlized), *poag* (post-aspirated & glottalized); **Phon** (phonation): *vl* (voiceless), *vd* (voiced), *pvd* (pre-voiced voiceless), *b* (breathy voiced), *c* (creaky); **Init** (initiation): *p* (pulmonic egressive), *gi* (glottal ingressive [implosive]), *ge* (glottal egressive [ejective]), *vi* (velaric ingressive); **Pri** (primary articulation diacritics): *none* (no primary articulation diacritics), *ret* (retracted), *adv* (advanced), *mc* (mid-centralized), *lwd* (lowered), *rzd* (raised), *ap* (apical), *lam* (laminal), *dvb* (double vertical bar below), *dhb* (double horizontal bar below), *ola* (open left angle below), *d* (dental), *c* (centralized); **Sec** (secondary articulation diacritics): *none* (no secondary articulation diacritics), *w* (labialized), *j* (palatalized), *v* (velarized), *ph* (pharyngealized), *jw* (palatalized and labialized), *jv* (palatalized and velarized), *wv* (labialized and velarized), *wph* (labialized and pharyngeal-

ized), *l* (lateral/lateral release), *r* (rhoticized/rhotic release), *wr* (labialized with rhotic release), *jr* (palatalized with rhotic release), *lv* (lateral and velarized), *lj* (lateral and palatalized), *lw* (lateral and labialized), *lph* (lateral and pharyngealized), *nas* (nasalized), *glt* (glottalized [vowels only]), *vs* (velar stop [clicks only]), *vf* (velar frication [clicks only]), *va* (velar affrication [clicks only]), *us* (uvular stop [clicks only]), *uf* (uvular frication [clicks only]), *ua* (uvular affrication [clicks only]); **Pros** (prosodic properties): *none* (no prosodic properties), *syl* (syllabic), *nsyl* (non-syllabic), *brev* (brief/breve), *long* (long/geminate), *ds* (downstep); **Tone** (tone markers): Chao digits 1-5 (Chao, 1968).

The three figures below show the inventory of English as given by the *Phoible* database in the *Phoible* feature system (Figure 1) and in the *Fonetikode* system (Figure 2), as well as the English inventory as given in the *Ruhlen* database by the *Fonetikode* system (Figure 3).

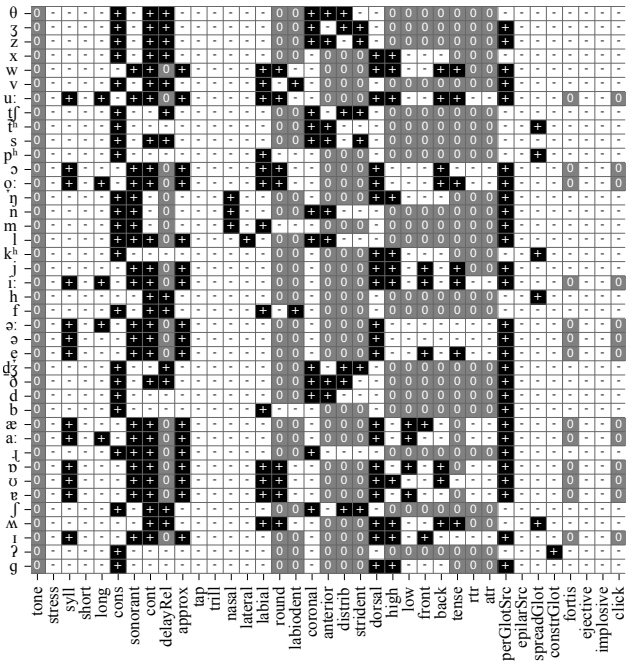

Figure 1: The inventory of English as given by *Phoible* and the *Phoible* feature system. Vertical axis: the segments in IPA notation; horizontal axis: the (abbreviated) feature names. Cell color is white for "–", gray for "0" and black for "+".

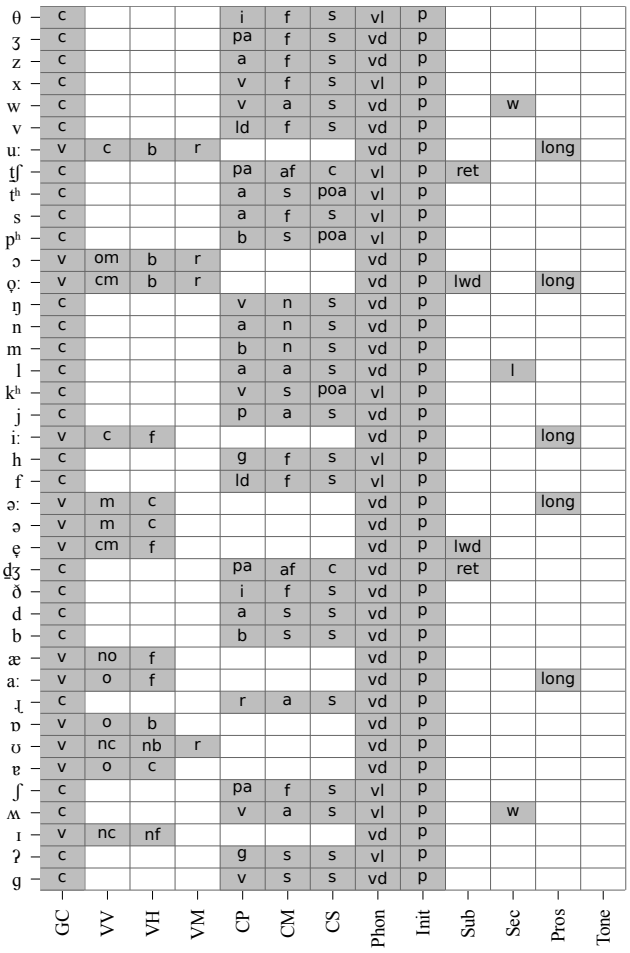

Figure 2: The inventory of English as given by *Phoible* (the same segments as in Figure 1) and the *Fonetikon* feature system. Vertical axis: the segments in IPA notation; horizontal axis: the feature names. Cell color is white for absent and "none" values, and gray for the other values.

2.3. Matching Languages: UULIDs

A very important issue when working with language-level databases in general, and exacerbated when combining several such databases, concerns the unique and reproductible identification of the entities involved (languages, language families, dialects, etc). In this paper we use an approach described in detail elsewhere (`https://github.com/ddediu/lgfam-newick/blob/master/paper/family-trees-with-brlength.pdf`) which combines four widely-used schemes for identifying linguistic entities: the ISO 639-3 three-letter codes (`http://www-01.sil.org/iso639-3`), the WALS three-letter codes (`http://wals.info/languoid`), the AUTOTYP numeric codes (`http://www.autotyp.uzh.ch/theory.html`), and the Glottolog alpha-numeric codes (`http://glottolog.org/glottolog/glottologinformation`).

For each language in the *Phoible* and *Ruhlen* databases, we have identified their codes in the four schemes and we have further created a unique combination of these that uniquely identifies a given linguistic entity across these four schemes. The format of this combination code, denoted here as *Universally Unique Language IDentifier* (UULID), is `[i-I][w-W][a-A][g-G]` where "I", "W", "A" and "G" stand for the ISO 639-3, WALS, AUTOTYP and Glottolog code(s) of the language, respectively. Please note that any of these codes can be missing (but not all of them at once) if the language is not defined in the corresponding scheme, or there can be

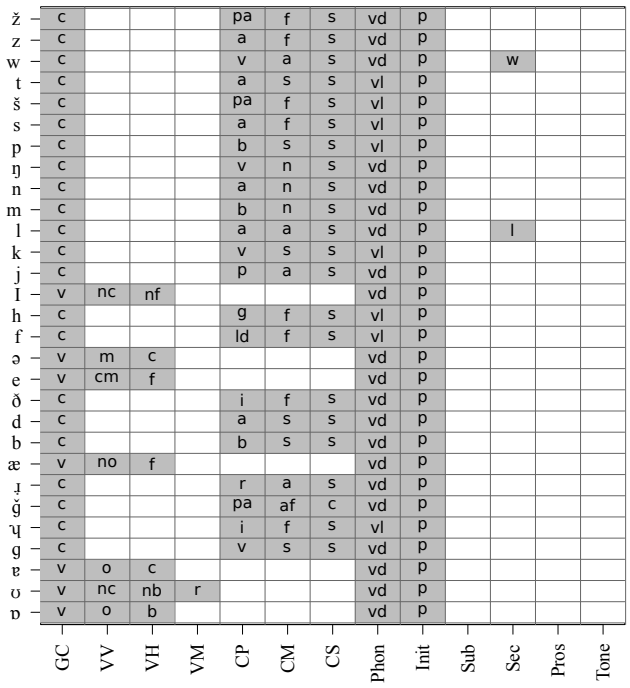

Figure 3: The inventory of English as given by *Rhulen* (different segments from Figures 1 and 2) and *Fonetikon* feature system (same as in Figure 2). Vertical axis: the segments in Ruhlen's notation (*NB.* we have mapped all of the APA and non-standard or atypical glyphs that appear in *Ruhlen* to their IPA equivalents, but we show here the original glyphs); horizontal axis: the feature names. Cell color is white for absent and "none" values, and gray for the other values.

more than one value (separated by "-") if, for example, a scheme distinguishes between several subdivisions that other schemes lump together. Some examples are (WALS language names, the Indo-European family): "German (Zurich)" `[i-gsw][w-gzu][a-1305-1306-1307][g-swis1247]`, "Urdu" `[i-urd][w-urd][a-2671][g-urdu1245]`, "Romani (Sepecides)" `[i-][w-rse][a-][g-]`.

2.4. Defining classes of segments

The two feature systems, *Phoible* and *Fonetikode*, were used to define *classes of segments*: both *Phoible* and *Fonetikode* were applied to the *Phoible* database, but only *Fonetikode* to the *Ruheln* database (because the *Phoible* system does not cover certain segments present there), resulting thus in *three* combinations of a segment database and feature system, denoted in the following *PP* (*Phoible* + *Phoible*), *PF* (*Phoible* + *Fonetikode*), and *RF* (*Ruhlen* + *Fonetikode*). Currently, there are 175 defined classes such as "segment", "mid vowel", "retroflex stop", "bilabial fricative" and "level tone", but the *Phoible* feature system cannot easily describe some of these (e.g., "level tone", "contour tone", "doubly articulated consonant").

Such classes of segments can be easily defined using our implementation in `R` (R Core Team, 2014), which is a very flexible and powerful system for specifying sets of features, their values, and combinations thereof. The implementation supports two ways of describing feature values: either as "+", "–" or "0" preceding the feature name (the convention used by the *Phoible* feature system), or as the feature name followed by ":" and the comma-separated feature values (the convention used by the *Fonetikode* feature systems). For example, the definition of a *segment* in the first system is "0tone" and in the second is "GC:c,v", while a *vowel* is defined as ("+syll & –cons & +sonorant") and as "GC:v", respectively. These defintions can be combined using boolean logic (`and`, `or`, `not`) and more advanced programming constructions, resulting in a system that has the expressive power of `R`.

More precisely, the fundamental function is

```
`.` <- function(phonemes, featvals)
```

where the `phonemes` is a `data.frame` representing the segment inventory of a language in a given feature system such that each segment ('phoneme') is represented by one row and each column gives the value for a feature of the system (e.g., for English in the *PF* case there would be 40 rows and 14 relevant columns including the actual segment, such that the first five rows would be as in Table 1). The `featvals` represents the description of the relevant combination of features and values that (partly) describe the class; for example, in the *PF* system, the segments can de defined as `.(phonemes, featvals="GC:c,v")`[7]. The `.()` returns a logical vector of the same length as the number of rows (segments) in `phonemes` with `TRUE` signalling that the segment fits the description in `featvals`, allowing the combination of multiple calls to `.()` to be combined using the logical operators `!`, `&` and `|` or any other processing on vectors of logical values. For example, the retroflexes can be defined as `.(phonemes,"GC:c") & .(phonemes,"CP:r")` using the logical conjunction of two subconditions (being a consonant and having a retroflex place of articulation). This allows the definition of a class to be encapsulated in a function, for example being a vowel can be defined as `vowels <- function(phonemes){ .(phonemes,"GC:v")}` allowing other definitions to reuse them, such as defining diphtongs as `diphtongs <- function(phonemes){ vowels(phonemes) & vapply(unique.feat.vals(phonemes, "VV"), function(s) length(s)==2, logical(1))}` (where `unique.feat.vals(phonemes, feature)` returns the unique values of the `feature` in the inventory of `phonemes`).

Full details and the actual implementation can be found in the GitHub repository `https://github.com/ddediu/phon-class-counts`.

[7] Of course, when using the `Phoible` feature system these descriptions are different; in this case `.(phonemes,"0tone")`.

Phoneme	GC	VV	VH	VM	CP	CM	CS	Phon	Init	Pri	Sec	Pros	Tone
ʔ	c				g	s	s	vl	p	none	none	none	none
ʃ	c				pa	f	s	vl	p	none	none	none	none
ɻ	c				r	a	s	vd	p	none	none	none	none
g	c				v	s	s	vd	p	none	none	none	none
ʍ	c				v	a	s	vl	p	none	none	none	none
...													

Table 1: The format of the `phonemes` argument to the '`.`' function exemplified by English in the *PF* case (showing the first 5 rows/segments).

3. Results

The two databases *Phoible* and *Ruhlen* contain distinct but overlapping sets of languages (Figure 4).

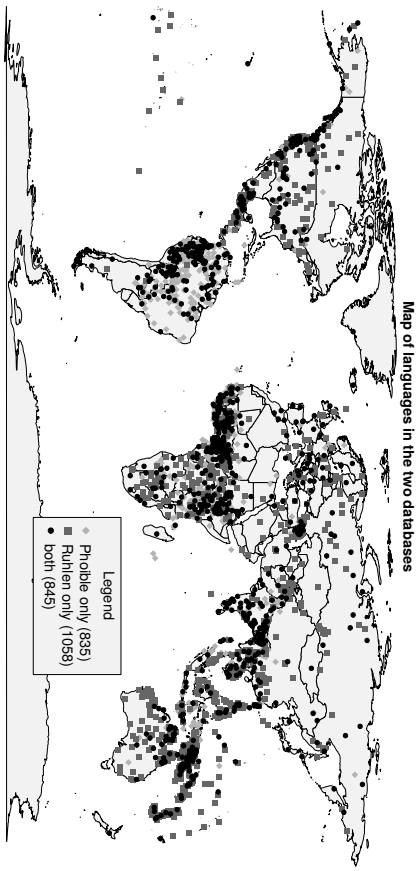

Figure 4: Map showing the databases' coverage.

For each of the three combinations of segment database and feature system (see Section 2.4. for details) *PP*, *PF*, and *RF*, there are currently 175 such classes defined, as follows (".v" stands for vowels, ".c" stands for consonants; for full details see `https://github.com/ddediu/phon-class-counts`): *segments*, *vowels*, *monophtongs*, *diphtongs*, *triphtongs*, *heights.v*, *lengths.v*, *long.v*, *nasal.v*, *round.v*, *high.v*, *mid.v*, *low.v*, *front.v*, *back.v*, *tense.v*, *lax.v*, *atr.v*, *rtr.v*, *raised.v*, *retracted.v*, *fronted.v*, *glottalized.v*, *laryngealized.v*, *unique.v*, *unique.nasal.v*, *heights_mono.v*, *heights_di.v*, *heights_tri.v*, *lengths_mono.v*, *lengths_di.v*, *lengths_tri.v*, *long_mono.v*, *long_di.v*, *long_tri.v*, *nasal_mono.v*, *nasal_di.v*, *nasal_tri.v*, *round_mono.v*, *round_di.v*, *round_tri.v*, *high_mono.v*, *high_di.v*, *high_tri.v*, *mid_mono.v*, *mid_di.v*, *mid_tri.v*, *low_mono.v*, *low_di.v*, *low_tri.v*, *front_mono.v*, *front_di.v*, *front_tri.v*, *back_mono.v*, *back_di.v*, *back_tri.v*, *tense_mono.v*, *tense_di.v*, *tense_tri.v*, *lax_mono.v*, *lax_di.v*, *lax_tri.v*, *atr_mono.v*, *atr_di.v*, *atr_tri.v*, *rtr_mono.v*, *rtr_di.v*, *rtr_tri.v*, *raised_mono.v*, *raised_di.v*, *raised_tri.v*, *retracted_mono.v*, *retracted_di.v*, *retracted_tri.v*, *fronted_mono.v*, *fronted_di.v*, *fronted_tri.v*, *glottalized_mono.v*, *glottalized_di.v*, *glottalized_tri.v*, *laryngealized_mono.v*, *laryngealized_di.v*, *laryngealized_tri.v*, *unique_mono.v*, *unique_di.v*, *unique_tri.v*, *unique.nasal_mono.v*, *unique.nasal_di.v*, *unique.nasal_tri.v*, *consonants*, *places.c*, *bilabial.c*, *labiodental.c*, *double_articulated.c*, *dental.c*, *alveolar.c*, *dental_alveolar.c*, *palatoalveolar.c*, *alveolopalatal.c*, *postalveolar.c*, *true_retroflex.c*, *palatal.c*, *velar.c*, *uvular.c*, *pharyngeal_epiglottal.c*, *glottal.c*, *labial.c*, *coronal.c*, *dorsal.c*, *guttural.c*, *manners.c*, *obstruent.c*, *voiced_obstruent.c*, *voiceless_obstruent.c*, *aspirated_obstruent.c*, *glottalized_obstruent.c*, *stop.c*, *voiced_stop.c*, *voiceless_stop.c*, *aspirated_stop.c*, *glottalized_stop.c*, *fricative.c*, *voiced_fricative.c*, *voiceless_fricative.c*, *affricate.c*, *sonorant.c*, *voiced_sonorant.c*, *voiceless_sonorant.c*, *glottalized_resonant.c*, *nasal.c*, *approximant.c*, *tapflap.c*, *trill.c*, *trill_tap.c*, *coronal_trill_tap.c*, *second_articulation.c*, *glottalized.c*, *uvt.c*, *uvt.stops.c*, *uvt.fricatives.c*, *uvt.affricates.c*, *uvt.nasals.c*, *uvt.approximants.c*, *lvt.c*, *lvt.stops.c*, *lvt.fricatives.c*, *lvt.affricates.c*, *lvt.nasals.c*, *lvt.approximants.c*, *ratio.voiced.voiceless.obstruents*, *ratio.voiced.voiceless.stops*, *ratio.obstruents.sonorants*, *egressive*, *implosive*, *ejective*, *click*, *voiceless*, *voiced*, *breathy*, *creaky*, *tones*, *level_tones*, *contour_tones*, *bilabial.fricatives*, *labiodental.fricatives*, *alveolar.fricatives*, *nonsibilant.dental.fricatives*, *sibilant.dental.fricatives*, *bilabiallabiodental.affricates*, *bilabial.affricates*, *retroflex.stops*, *retroflex.fricatives*, *retroflex.affricates*, *retroflex.nasals*, *retroflex.approximants*.

For any given such class (say, *segments* or *vowels*) and language (say, English) in a given database and feature system (say, *PP*), the system automatically selects the language's segments that belong to this class and computes their count. Thus, for English, there are 40 seg-

ments [g ʔ ʃ ɪ ʊ ɐ ɒ ɾ ʍ aː æ b d ð d͡ʒ e̞ ə ɔː f h iː j kʰ l m n ŋ o̞ː ɔ pʰ s tʰ t͡ʃ uː v w x z ʒ θ] in *PP*, 40 segments [ʔ ʃ ɾ g ʍ ɐ ɒ ɪ ʊ aː æ b d ð d͡ʒ e̞ ə ɔː f h iː j kʰ l m n ŋ o̞ː ɔ pʰ s tʰ t͡ʃ uː v w x z ʒ θ] in *PF*, and 29 segments [ɥ g ɐ ɒ ʊ ɡ̆ ɻ æ b d ð e ə f h I j k l m n ŋ p s š t w z ž] in *RF*; and for the same language, there are 13 vowels [ɪ ʊ ɐ ɒ aː æ e̞ ə ɔː iː o̞ː ɔ uː] in *PP*, 13 vowels [ɐ ɒ ɪ ʊ aː æ e̞ ə ɔː iː o̞ː ɔ uː] in *PF*, and 7 vowels [ɐ ɒ ʊ æ e ə I] in *RF*. Therefore, for each of the three combinations of database and feature system (*PP*, *PF* and *RF*) we computed for each language and class the number of segments in the class (as well as the actual segments that belong to the class). Given that some classes cannot be estimated for the *Phoible* feature system, we ended up with 167 classes in 1680 languages for *PP* (classes *laryngealized.v*, *laryngealized_mono.v*, *laryngealized_di.v*, *laryngealized_tri.v*, *double_articulated.c*, *second_articulation.c*, *level_tones*, *contour_tones* are not defined), 175 classes in 1680 languages for *PF*, and 175 classes in 1903 languages for *RF*.

3.1. Correlation Phoible – Fonetikode

What is the relationship between the two feature systems we are using, *Phoible* and *Fonetikode*? Given that the two systems are both applied on the *Phoible* database, we compared the class counts and composition generated by these systems. The Pearson correlations between the counts are all highly significant ($p < 0.01$ after Bonferroni correction) and vary between 0.16 and 1.00 with an average of 0.88 (and a median of 0.95). The Jaccard index[8] of the actual segments in the classes averaged across languages varies between 0.00 and 1.00, with a mean of 0.92 and median 0.99. Therefore, for the vast majority of classes the two feature systems strongly agree producing very similar classes.

3.2. Correlation Phoible – Ruhlen

What is the relationship between the two databases, *Phoible* and *Ruhlen*? Given that we applied the *Fonetikode* feature system to both databases, we compared the classes counts and composition on the 845 languages that are present in both databases. The Pearson correlations between the counts are all highly significant ($p < 0.01$ after Bonferroni correction) and vary between -0.03 and 0.94 with an average of 0.51 (and a median of 0.57). The Jaccard index of the actual segments in the classes averaged across languages varies between 0.24 and 1.00, with a mean of 0.77 and median 0.82. Therefore, the two databases tend to produce similar (but not identical) classes of segments for the languages shared between them.

[8] A measure of similarity between two sets A and B defined as $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, i.e. the ratio between the size of the intersection and the size of the union of the two sets, varying between 0.0 (nothing in common) to 1.0 (identical).

3.3. Relations Between Classes

To investigate the relationships between the classes, we conducted Principal Component Analysis (Jolliffe, 2002) separately on the three database and feature system cases. More precisely, given a database and feature system (for example, *PP*), we constructed the matrix of all languages × all classes where each cell (l, v) contains the number of segments class v has in language l, excluding the cells with missing data (in the example, a 1680×159 matrix), and we applied PCA on this matrix considering the languages as observations (rows) and the classes of segments as variables (columns).

For *PP*, the first three Princial Components (PCs) explain 40.6% of the variance (with 90% of the variance being explaind by the first 37 PCs): PC_1 (18.1%) expresses the agreement among all features (almost all tend to have loadings of the same sign), PC_2 (12.9%) distinguishes consonants from vowels, while PC_3 (9.6%) is hard to interpret.

For *PF*, the first three Princial Components (PCs) explain 38.9% of the variance (with 90% of the variance being explaind by the first 40 PCs): PC_1 (16.8%) expresses the agreement among all features, PC_2 (12.0%) distinguishes consonants from vowels, while PC_3 (10.2%) is hard to interpret. These results are quite similar to *PP*, which is to be expected given the high correlations between the two feature systems.

For *RF*, the first three Princial Components (PCs) explain 39.6% of the variance (with 90% of the variance being explaind by the first 33 PCs): PC_1 (17.0%) expresss the agreement among all features, PC_2 (15.1%) distinguishes consonants from vowels, while PC_3 (7.5%) is hard to interpret.

Thus, in all three cases, the most important tendency ($\approx$ 17% of the variance) is shared among all classes, followed by the rough distinction between vowels and consonants ($\approx$ 13% of the variance).

3.4. Cross-linguistic Distribution

Figures 5 and 6 show the distribution of the number of segments and of the retroflex stops as given by *Phoible* with the *Fonetikode* feature system (*PF*), but similar maps can be drawn for each class in each database and feature system.

What can be immediately seen is that there is a lot of diversity with interesting geographic patterns that require further analysis.

4. Discussion and conclusions

The system introduced in this paper is freely available in the GitHub repository `https://github.com/ddediu/phon-class-counts` and allows the flexible definition of classes of segments given a feature system that describes each segment in terms of feature values. Thefore, the system is easily extensible to new feature systems, new classes of segments and to new segment-level databases. Here we applied it to the two segment-level databases *Phoible* (Moran et al., 2014) and *Ruhlen* (Creanza et al., 2015) using two features systems

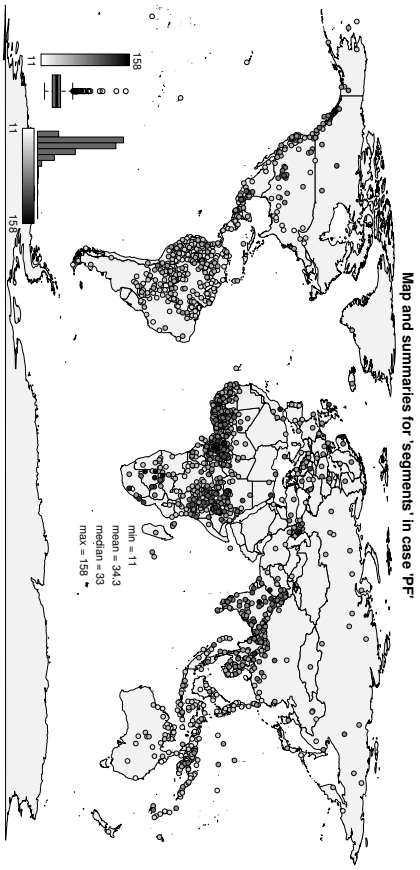

Figure 5: Map showing the number of *segments* per language in the *PF* case. Each circle is a language with color represeting the number (white = minimum to black = maximum). The bottom-left boxplot and histogram show the distribution of the counts, and the text right of Africa gives basic summaries. The higher counts are visually stacked above the lower counts making them more visible when mutiple languages overlap.

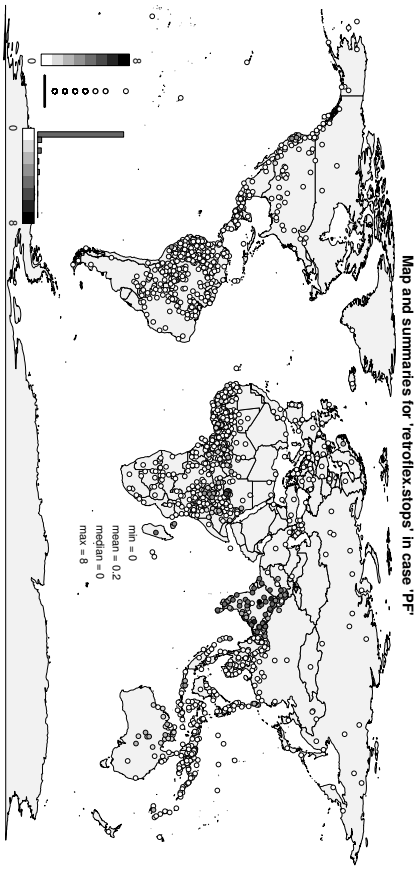

Figure 6: Map showing the number of *retroflex stops* per language in the *PF* case; the conventions are as in Figure 5.

(*Phoible* and our own IPA-inspired *Fonetikode*) resulting in higher-level descriptions of the languages in terms of classes such as "vowel", "retroflex stop" or "click". Such abstract descriptions can then be used to explore the patterns of cross-linguistic diversity, their causes and their relationship to extra-linguistic factors and processes including historical, economic and climatic, among others. We hope that the system introduced and the results reported here will be useful to a wide range of scientists adressing issues related to the pattens of cross-linguistic diversity.

5. Acknowledgements

We wish to thank Steve Moran for help with the *Phoible* database, and to the creators and maintainers of LaTeX, `R` (R Core Team, 2014), `Sweave` (Leisch, 2002) and `RStudio` (RStudio Team, 2015). This research was funded by the *Netherlands Organisation for Scientific Research* (NWO) VIDI grant number 276-70-022 to DD.

6. Bibliographical References

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press: Berkeley, California.

Creanza, N., Ruhlen, M., Pemberton, T. J., Rosenberg, N. A., Feldman, M. W., and Ramachandran, S. (2015). A comparison of worldwide phonemic and genetic variation in human populations. *PNAS*, page 201424033.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info.

Martin Haspelmath, et al., editors. (2005). *The World Atlas of Language Structures*. Oxford University Press.

Hayes, B. (2009). *Introductory Phonology*. Blackwell.

International Phonetic Association. (2005). International phonetic alphabet. Technical report, International Phonetic Association. http://www.arts.gla.ac.uk/IPA.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer Verlag: NY, 2 edition.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle et al., editors, *Compstat*

2002 — Proceedings in Computational Statistics, pages 575–580. Physica Verlag, Heidelberg. http://www.stat.uni-muenchen.de/ leisch/Sweave.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge studies in speech science and communication. Cambridge University Press, Cambridge.

Moisik, S. R. and Esling, J. H. (2011). The 'whole larynx' approach to laryngeal features. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS XVII)*, pages 1406–1409.

Steven Moran, et al., editors. (2014). *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://phoible.org.

R Core Team, (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

RStudio Team, (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. http://www.rstudio.com/.