Phonetics Information Base and Lexicon

Steven Paul Moran

Department of Linguistics, University of Washington*

Summary

In my dissertation I answer the question of whether more sophisticated, knowledge-based approaches to data modeling, coupled with a broad cross-linguistic data set, can extend previous typological observations and provide novel ways of querying data about the sound systems of the world's languages. The model that I implement facilitates testing typological observations by aligning data models to questions that typologists wish to ask. The technological infrastructure that I create is conducive to data sharing, extensibility and reproducibility of results. I use the data set and data models in this work to validate and extend previous typological observations.

In doing so, I revisit the typological facts proposed in the linguistics literature about the size, shape and composition of segment inventories in the world's languages and find that they remain similar even with a much larger sample of languages. I also show that as the number of segment inventories increases, the number of distinct segments also continues to increase. And when vowel systems grow beyond the basic cardinal vowels, they do so first by length and nasalization, and then diphthongization.

Moving beyond segments, I show that distinctive feature sets in general lack the typological representation needed to straightforwardly map sets of features to the segment types found in a broad set of language descriptions. Therefore, I extend a distinctive feature set, devise a method to computationally encode features by combining feature vectors and assigning them to segment types, and create a system in which users can query by feature, by sets of features that define natural classes, or by omitting features in queries to utilize the underspecification of segments. I use this system and reinvestigate proposed descriptive universals about phonological systems and find that some, but not all universals hold up to the more rigorous testing made possible with this larger data set and a graph data model.

Lastly, I reevaluate one of the many purported correlations between a non-linguistic factor and language: the claim that there exists a relationship between population size and phoneme inventory size. I show that this finding is actually an artifact of a small data set, which constrains the use of more nuanced statistical approaches that can control for the genealogical relatedness of languages. Thus, in this work I illustrate how researchers can leverage the data set and data models that I have implemented to investigate different aspects of languages' phonological systems, including the possible impact of non-linguistic factors on phonology. After a brief overview of the background and challenges faced in my work, I describe the main outcomes of my research in the following sections of this extended dissertation abstract.

*Now in the General Linguistics Department at the University of Zurich and in the Quantitative Language Comparison Research Unit at the University of Munich.

1 Overview

Phonological typology typically involves comparing languages by the number or types of sounds, or *segments* when encoded by graphic symbols, that they contain. My work draws on linguistic research in segmental phonology and distinctive feature theory, and on computational research in data modeling and knowledge representation. My colleagues and I have created a cross-linguistic data set and I have modeled this data set in ways that allow researchers to investigate the variation of phonological systems across languages at the level of segments and at the level of distinctive features – the phonetic 'atoms' that are combined to articulate the sounds found in any human language.

The motivation behind this work was to collect a much larger and broader cross-linguistic sample of phonological inventories than what was previously available and to model the data in an interoperable way so that users could federate disparate linguistic and non-linguistic information and pose novel questions on the combined data set. I call this resource the Phonetics Information Base and Lexicon (PHOIBLE).[1]

PHOIBLE incorporates the segment inventories from the Stanford Phonology Archive (SPA; Crothers et al. 1979), the UCLA Phonological Segment Inventory Database (UPSID; Maddieson 1984, Maddieson & Precoda 1990) and the *Systèmes alphabátiques des langues africaines* (AA; Hartell 1993, Chanard 2006). Additionally, the genealogical and geographical coverage of these combined sound system inventories is expanded by the work that my colleagues and I undertook in extracting phonological inventory data from hundreds of grammars and phonological descriptions. The combined PHOIBLE data sample contains 1336 segment inventories, which represent 1089 distinct languages, or roughly 16% of the world's estimated 6909 languages, as listed in the Ethnologue (Lewis 2009). Inventories range in detail from phonemic descriptions to fuller phonological descriptions including phonemes, allophones, their conditioning environments and additional information like phonological rules and a description of marginal sounds.

From the beginning my goal has been to create a tool for typology that is extensible and that can also interoperate with additional linguistic and non-linguistic data sets. Although the inventories in PHOIBLE represent a convenience sample, i.e. a set of languages chosen from sources that are readily available,[2] each segment inventory is associated with data regarding its genealogical affiliation, including its language family stock from the Ethnologue (Lewis 2009) via Multitree (LINGUIST List 2009) and its language genus from the World Atlas of Language Structures (WALS; Haspelmath et al. 2008). Geographical information for each language also comes from the Ethnologue (country and geographic region) and WALS (geo-coordinates). Genealogical and geographic information is pertinent to statistical sampling in linguistic typology so that factors of shared descent and areal diffusion can be accounted for and can be used to inform statistical observations. Non-linguistic information, such as demographic data, is also included so that various cross-cultural and cross-disciplinary studies can be undertaken.

A major challenge in this work has been addressing the question of how to bring together these segment inventory databases, which are heterogeneous in format, encoding and content, into an accessible data model that is extensible and which can integrate additional linguistic and non-linguistic information. Before the integration processes and the resulting data models could be instantiated, however, there were many methodological considerations at the linguistic and technological levels that had to be identified and addressed, which I do in Chapter 2: Background.

[1] `http://phoible.org/`

[2] Note that around half of the world's languages have not been described in any detail.

2 Background

I begin this chapter by defining the conventions and terminology used in my work. Of particular importance is the disambiguation of terminology that intersects at linguistics and technology, especially in the domains of writing systems and character encodings, e.g. 'character', 'grapheme' and 'script'. I then provide a brief description in Section 2.2 of the fundamental linguistic theories pertinent to my research: segmental phonology and distinctive feature theory.

Segmental phonology is the study of speech sounds modeled as abstract segments that are discrete and serially ordered. Its practitioners investigate the distribution of sounds and their patterning by means of a theoretical framework that strives to answer questions regarding the nature of phonetic alternations and contrastive sounds that trigger lexical or grammatical differences in languages. Each spoken language can be described with a language variety-specific set of segments, which it uses to form and differentiate words.

Distinctive feature theory is born out of work in segmental phonology and is considered one of the most important contributions to linguistics in the 20th century because of the explanatory power that it provides. It has a long tradition in linguistics, in such works as Trubetzkoy 1939, Jakobson 1949, Jakobson et al 1952 and Jakobson & Halle 1956. By building on the work of members of the Prague Linguistic Circle (or Prague School) and the American structuralists in the early to mid 20th century, Noam Chomsky and Morris Halle created generative phonology, the dominant phonological theory for the next 30 years or so (Chomsky & Halle 1968). In generative grammar, phonological representations were modeled as sequences of segments composed of distinctive features. This provided a framework for phonologists to describe phonological rules and derivations, and levels of phonological representations through fully explicit algorithms using linear sequences of matrices of distinctive feature values.

Distinctive features represent abstract properties of speech sounds, typically modeled on phonetic correlates rooted in human anatomy. The mental representation of a speech sound was originally modeled as an unorganized set of feature values. Two speech sounds contrast if they differ by at least one distinctive feature. Jakobson's approach was to keep the number of distinctive features at a minimum (e.g. Jakobson 1949). For example, an eight vowel system requires 28 binary relations if each vowel opposes every other vowel. These 28 binary oppositions can be expressed in terms of three distinctive features (e.g. [high], [back] and [round] in SPE), resulting in only three oppositions. This approach reduces entropy, so that there is less functional load involved in the storing and processing of language for the speaker and listener. Indeed much of this work was inspired by Shannon's work on information theory.

With the vocabulary in place and the basic linguistic theories explained, in Section 2.3 I discuss the linguistic and technological challenges involved in developing a cross-linguistic data set to compare and characterize the distribution of linguistic phenomena. Although my focus is on data from segmental phonology and distinctive feature theory, the broader challenges that I faced are applicable to developers of other typological databases. A theoretical issue is whether typology can be undertaken with language-specific analyses or if separate over-arching cross-linguistic comparative concepts are needed. This is an area of an ongoing heated debate, cf. Lazard 2006, Haspelmath 2007, 2010, Newmeyer 2010, Bickel 2010. The problem is exemplified by the types of claims made by researchers using typological databases, which can bring together a wide range of different descriptions of languages.

Large samples of diverse data also raise the issue of how statistical sampling should be used to account for the various types of bias that are inherent in linguistic data sets. The challenge of deriving a cross-linguistic language sample that captures genealogical, areal and typological diversity was raised as early as Sherman 1975. Later, statistical methods based on classical sampling theory were described as not tenable for most

typological data (Janssen et al. 2006). The foundation of many of these methods requires a population from which a random sample can be drawn and one that fits a normal distribution. However, language data are a skewed population of data points due to factors including the diffusion of typological features through shared descent and geographic proximity. Of course one can draw a random sample from the population, but it might not be representative for the question being asked. Thus, the question of how to establish an ideal sample for purposes of statistical evaluation is central to typological methodology. In the section on sampling, I describe the statistical biases and I give a historic overview of statistical sampling techniques that have been proposed to address these confounding factors.

Another problem related to typological comparison involves the analysis of data; the problem is captured by the paradox of using linguistic theory to document and describe languages, but the need to abstract away from theory to undertake cross-linguistic comparison (Hyman 2008). Keeping track of different analyses from different authors is also an issue of data provenance. New analyses may involve the reinterpretation of older analyses, particularly when one wants to standardize across descriptions to create comparative concepts. Lastly, the practical implementation of a cross-linguistic data set to undertake phonological typology requires the standardization of segments at both the linguistic and technological levels. Once all of these issues have been addressed, the next question involves asking what type of questions can be asked of the data set given the model(s) in which the data are encoded.

3 Data Modeling

There are many ways to model data. Some methods are well researched, considered mature and are used in all kinds of applications across many different industries. Other methods represent the state-of-the-art in data structures and algorithm design and are being researched and developed at the peripheries of computer science. While there are many different ways to think about and model data, different methods have different strengths and weaknesses for different purposes. Therefore it is necessary to model different data types with appropriate data structures to enable the desired questions to be answered. In Chapter 3, I contrast three different ways of modeling data and I describe in detail knowledge representation in computational theory and how it can be used to query the PHOIBLE data set from different perspectives.

Section 3.1 provides a brief introduction to different data models and some examples with linguistic data. In general it is important that data are easily interpretable (Bird & Simons 2003, Abney & Bird 2010); a simple machine readable storage model is a practical way to make data available to a large audience.[3] Thus, flat file tables are one format in which the PHOIBLE data set is made available. The tables are also convenient as an input format for statistical packages and programming scripts, as I show in Chapters 5 & 7, in which I investigate various properties of segment inventories and a reported correlation between segment inventory size and population size.

In Section 3.2 I describe in detail the three PHOIBLE data models: flat file tables, a relational database and an RDF graph. I provide many examples of how a user might query each type and illustrate the strengths and weaknesses of the underlying data model. In Section 3.3 I discuss aspects of knowledge representation and how formal logic constraints can be integrated with the PHOIBLE RDF graph to create a 'knowledge base', i.e. a collection of assertions about phonological inventories and data related to those languages in a formal knowledge representation language. The graph model coupled with a knowledge representation

[3]This is a particularly important point in regard to linguistic data because the object of study – human languages – are disappearing at an alarming rate.

formalism allows researchers to manipulate aspects of the PHOIBLE data set, such as specifying that the distinction between long and short vowels should be collapsed or that diphthongs should be ignored in a query, without changing the underlying data and thus allowing the researcher to apply his or her own analytical preferences to the data. Additionally, I have defined an ontology to encode concepts and their relationships in the data, so that a vocabulary of phonetic features has been given hierarchical structure to represent feature geometries, which can then be used to query the PHOIBLE data set or selected portions of it. Users can extend this ontology or define their own ontologies to interact with the data in PHOIBLE in different ways. By merging the PHOIBLE segment inventory and distinctive feature graphs, I provide researchers with a novel tool to investigate data about the world's sound systems at the level of articulatory features. I use this tool in Chapter 6 to test claims about universal properties of phonological systems.

As I show in my work, there is no one-data-model-fits-all approach for investigating questions in phonological typology. Data are ideally modeled in ways that are flexible such that different typological observations can be tested in appropriate ways and the same questions can be approached from multiple perspectives. First, however, the data must be brought together into one linguistically and technologically interoperable data set. This is the topic of Chapter 4.

4 PHOIBLE

In Chapter 4 I explain how syntactic and semantic interoperability are achieved by extracting the segment inventory data from various disparate formats, bringing the data together into one data set that adheres to a well-defined standard of segments and their distinctive features, and then modeling the data set into formal data models that are aligned to questions that linguists wish to ask. I begin by describing the resource that I have developed.

PHOIBLE is an online repository of cross-linguistic phonological segment inventory data that contains additional linguistic and non-linguistic information about languages. All segment data in PHOIBLE were standardized and compiled into a single data repository through a process commonly called Extract, Transform and Load (ETL; Inmon 1992, Kimball 1996). The SPA, UPSID, AA and PHOIBLE inventories each underwent an individualized ETL process because each data source provided its own set of challenges in forming a unified, all-Unicode IPA data repository. Additional linguistic and non-linguistic data were added to the PHOIBLE database via tables and associated with segment inventories via their ISO 639-3 codes, so that these data can be easily updated in future releases.

I then describe the challenges and the ETL processes used to bring together SPA, UPSID_{451}, AA and the PHOIBLE inventories into one interoperable data set. The initial lack of interoperability between the resources described in Chapter 4 highlights the need for technological infrastructure that supports research across disparate data sets. The use of standards such as ISO 639-3, IPA and Unicode will promote interoperability between PHOIBLE and other resources, such as those being added to the Linguistics Linked Open Data cloud, an effort being spearheaded by the Working Group on Open Data in Linguistics.

Combining the SPA, UPSID_{451}, AA and PHOIBLE segment inventories together results in a sample that represents 16% of the world's languages. There is no simple and straight forward means to evaluate the genealogical coverage of a large typological data sample on a family-per-family (or genus-per-genus) basis. Even though many genealogical language classifications are working hypotheses, it is nevertheless important to establish what the genealogical coverage of a typological data set is, thereby allowing the coverage of different data sets to be compared. In Section 3.4 I describe a method I developed for evaluating the

genealogical coverage of a data set by using a list of ISO 639-3 language name identifiers and simple XML representations that represent language family trees, extracted from the Linguist List's Multitree project. I use this method to assess the genealogical coverage of PHOIBLE by comparing its contents with language families in the Ethnologue 15th edition, currently the most-up-to-date data available through Multitree.

5 Segments and Inventories

In my work I have taken a data-driven approach in collecting segment inventories from different tertiary databases and from secondary resources like grammars and phonological descriptions. These resources vary widely in their descriptions and analyses of languages' segment inventories. The technological architecture that I have developed allows users to decide whether they want to keep or remove certain segment types from their experiments, such as diphthongs, tone or vowels with secondary phonation types. In Chapter 5, I investigate whether descriptive typological facts about segment inventories still hold up when we probe a much larger database of languages. In Section 5.2, I provide some background about the resources and work from which I examine properties of segment inventories.

In Section 5.3 on the distribution of segments, I use the denormalized table format of the PHOIBLE data set and load the data tables into statistical software to examine and illustrate properties of segment inventories and the distribution of segments cross-linguistically. I examine the distribution of segment types in PHOIBLE and show that as the number of segment inventories increases, the number of segment types seems to be increasing in a quadratic curve with no asymptote in sight. Or in other words, with each newly added language to PHOIBLE, there are a number of new sounds that do not occur in any other language in the sample. In this section I also investigate the frequency of segment types in PHOIBLE by devising and coding a resampling method that estimates genealogical bias. The resampling technique lets us infer the probable distribution of segment type frequencies by repeatedly sampling a random language representative from groups such as language families and systematically recomputing a statistical estimate by randomly sampling from subsets within a data set.

In Section 5.4 on segment inventories, I review some of the typological facts put forth by research undertaken with SPA (Crothers et al. 1979), UPSID_{317} (Maddieson 1984, 1986, Lindblom & Maddieson 1988), UPSID_{451} (Maddieson & Precoda 1990, Maddieson 1991) and WALS (Maddieson 2008a,c,b). I present these data within a historical perspective by comparing the SPA, UPSID_{451} and PHOIBLE inventories. I then apply the genealogical stratification method to the PHOIBLE data set and I evaluate the size and distribution of total segment inventories, and consonant, vowel and tone inventories separately. I show that Maddieson's findings about the sound systems of the world's languages generally still hold even as the size of a segment inventory databases increases. The mean and median figures from the genealogically stratified PHOIBLE sample are similar to those given by Maddieson through his work with the UPSID_{451}, UPSID_{317} and WALS samples.

Another topic of typological interest, particularly in the area of investigating language complexity in phonological systems, is the balance between consonants and vowels across inventories. This is the topic of Section 5.5. To summarize, the PHOIBLE data set shows a weak correlation between the number of consonants and vowels. There is no correlation between the number of consonants and tones in languages, nor is there a correlation between the number of vowels and tones.

In Section 5.6, I revisit Crothers's (1978) observation that vowel systems in most languages contain /i, a, u/. I look at relationships that hold among vowel systems in the PHOIBLE data set by creating distance

matrices using the Jaccard index, and preliminarily pointwise mutual information, and then visualizing these through multidimensional scaling. MDS visualizes some of the patterns that are inherent in the vowel space of inventories in PHOIBLE, e.g. vowel systems seem to grow by non-vowel quality distinctions like nasalization, lengthening and diphthongization. They then tend to pattern in front and back pairs. The smallest vowel systems tend to start with /i, a, u/.

Apart from describing the current state of what we know about the world's languages' phonological systems, another goal of my work is to provide novel access to phonological inventories and their associated data at a level deeper than the segment, that is, at the level of distinctive features. Chapter 6 is concerned with distinctive features and how to model them and use them to investigate phonological inventories at the sub-segment level.

6 Distinctive Features

In this chapter I set out to develop a mechanism for examining segment inventories at the level of distinctive features to investigate claims of descriptive universals in phonological systems. To do so, I begin with a brief discussion of segments and features in Section 6.2 and then I show in Section 6.3 that distinctive feature sets in general lack the typological representation needed to straightforwardly map each segment type in PHOIBLE to a set of features.

Therefore, in Section 6.4 I investigate the different types of segments and I outline how to compositionally encode features by combining feature vectors and assigning them to segment types. I developed a method for compositionally combining feature vectors to automatically derive features for segment types in the PHOIBLE data set that are not defined in Hayes 2009, which has the most comprehensive coverage of the IPA. I show that complex segments pose a serious challenge to automatic feature vector assignment because their component segments' feature vectors may not logically combine the way that segments and diacritics do. Contour segments also pose a challenge due to their temporal encoding of features changing through time. I proposed two ways of encoding contour segments for analysis.

The segment types and their features vectors are modeled in an RDF/OWL knowledge base, which provides the functionality for the user to query across segment inventories at the feature level. The user can query by feature, by sets of features that define natural classes, or by omitting features in queries to utilize the underspecification of segment types. The RDF/OWL model also provides structure that allows for the hierarchical organization of features into a feature geometry, which can be used to query inventories, and the model provides additional functionality to use logical operators and constraints in queries. My intent was to build a computational tool to allow researchers to undertake typological comparisons of segment inventories at the level of features. The system I have built does not rely on any particular feature set and the technologies I use allow users to plug other distinctive feature sets into the PHOIBLE architecture by mapping feature vectors to segment types, defining them in RDF, and merging the graphs.

To investigate segment inventories at the level of features, the approach I take is to combine two RDF graphs, namely the PHOIBLE segments and distinctive features graphs, into one combined RDF graph for querying. I use the SPARQL graph query language to investigate descriptive universals and provide examples for users who wish to probe the PHOIBLE knowledge base. In Section 6.5, I use the RDF/OWL knowledge base of PHOIBLE segment inventories and distinctive features from Hayes′ to revisit some of the universals of phonological inventories stated in Hyman 2008, a comprehensive study of universals in phonological systems, such as "all languages have coronals" and "every phonological system has at least one front vowel or the

palatal glide /j/". I show that at least one of these claims, namely that all phonological systems contain at least one coronal phoneme, does not hold on the extended PHOIBLE data set. There are other assumptions about segment inventories that are also important to test. However, I have not yet undertaken these studies. For example, one assumption is that languages with more fricatives will have a higher number of consonants overall. The data to address this question are easily attained from the PHOIBLE knowledge base by querying inventories for fricative and consonant segment types: for each inventory, get its number of fricatives and consonants by querying for all segments that are [−SONORANT], [+CONTINUANT] and [+CONSONANTAL], respectively. These queries would be difficult, or at least time-consuming, at the level of segments because a list of all fricative and consonant segment types in PHOIBLE would have to be identified and there are currently over 1700 segment types. Thus with PHOIBLE's inventories and the Hayes′ feature set, the technological infrastructure is in place for researchers to investigate many aspects of phonological systems and how they pattern. Going forward, certain claims are gaining attention nowadays that have to do with proposed correlations between certain aspects of phonological systems and non-linguistic factors. PHOIBLE is also an appropriate tool and data set to revisit these claims, as I show in Chapter 7.

7 Case Study: Phoneme Inventory Size and Population Size

Chapter 7 is a case study that uses the PHOIBLE data set to revisit the claim that there exists a correlation between population size and phoneme inventory size, as speculated in Haudricourt 1961 and Trudgill 1997, 2002, and empirically tested and reported in Hay & Bauer 2007. Hay & Bauer use a LOWESS statistical model with significance assessed with Spearman's rho on a set of 216 languages and find a positive correlation between population size and phoneme inventory size. My study addresses the shortcomings of this study by using a much larger data set with wider and deeper genealogical coverage and a hierarchical linear model to control for the genealogical relatedness of languages. I show that there is no correlation between population size and phoneme inventory size, once language family is accounted for.

My work also casts serious doubts on the results of studies that assume a positive correlation between population size and phoneme inventory size. For example, Atkinson (2011) proposes that a single language origin in Africa is supported by an out-of-Africa serial founder effect in which average phoneme inventory size decreases as one moves away from Africa. This analysis crucially depends on a correlation between population size and phoneme inventory size.

This case study shows how one might use PHOIBLE to investigate one of the many reported correlations between linguistic and non-linguistic factors, particularly claims regarding societal effects on language structure. I also discuss some of the methodological considerations in undertaking studies using statistical methods with phonological typological data and I illustrate how one might use PHOIBLE to investigate claims of correlations between non-linguistic factors and the phonological system, e.g. the claim that there exists a correlation between climate and the phonological system, i.e. languages spoken in warm climates use relatively more high-sonority sounds than those spoken in cold climates (Munroe et al. 1996, Munroe & Silander 1999, Fought et al. 2004, Ember & Ember 2007, Munroe et al. 2009).

8 Conclusion

My work contributes a large phonological typology data set to the field and makes these data openly available in different formats for researchers to use. These data are far from perfect, but they provide a new and richer

perspective on phonological systems of the world's languages. Coupled with additional linguistic and non-linguistic data, this data set provides a rich resource for undertaking phonological typology and it contains data pertinent to statistical sampling. My aim has been to model these data in formats that are extensible and interoperable, so that PHOIBLE can continue to grow and be integrated with new sources of data, such as lexicons, corpora, and non-linguistic data points like climate data and socio-economic variables like gross domestic product (GDP), etc.

In my dissertation I have raised and addressed several challenges pertinent to linguistics and the technological implementation of linguistic data, including:

- encoding linguistic segments in Unicode IPA for standardization and segment interoperability
- providing the Hayes 2009 distinctive feature set in Unicode and extending its incomplete IPA coverage as "Hayes Prime" that maps all unique Unicode characters to a vector of distinctive features; thus providing the basis for all segments types in PHOIBLE to receive a feature vector
- devising methods to automatically assign feature vectors to all segment types in inventories in PHOIBLE to achieve full typological coverage
- modeling PHOIBLE's data set in data structures that facilitate testing typological observations
- attaining structural interoperability of segments, segment inventories and distinctive features by modeling them in the RDF and OWL data models
- providing a feature geometry based on Hayes Prime and encoded in OWL

I have developed technological architecture that allows users to:

- query segment inventories at the level of segments and distinctive features
- query segment inventories by various linguistic and non-linguistic variables, e.g. segment class (i.e. consonant, vowel, tone, diphthong, etc.), language family or genus, geographical region, country or geo-coordinate, population, etc.
- access the data in various formats, including flat file tables, a relational database and an RDF graph model
- add information to the data set by using Linked Data
- manipulate the "surface" data set without changing its underlying contents by using OWL logic constructions and constraints on the RDF segments and features graphs
- test for correlations between linguistic and non-linguistic factors
- extract sample sets that adhere to genealogical and/or geographical constraints

Using the technological infrastructure and the data instantiated with it, including the segment inventories from three databases and the hundreds of additional inventories extracted from source documents, I revisit some of the typological facts put forth about segments and segment inventories in the world's languages. I show that:

- in general segment frequencies and the mean size of inventories remain close to the figures put forth in Maddieson 1984 and subsequent work using UPSID
- after taking into account genealogical skewing, segment types frequently found in most languages tend not to be far off from their frequency in the combined PHOIBLE data set, which is not genealogically balanced
- as segment inventories have been added to PHOIBLE, the number of new distinct segment types continues to increase at a rate that is not asymptotic
- there is a weak correlation between the number of consonants and vowels in segment inventories
- there is no correlation between the number of consonants or the number of vowels and tones in languages
- Crothers's (1978) observation that vowel systems typically have /i, a, u/ holds and I show with multidimensional scaling that vowel systems tend to expand beyond cardinal vowels by first adding a lengthened series of vowels, then a series of nasalized vowels, and then diphthongs

By building a system that allows researchers to query segment inventories at the level of distinctive features, I show that:

- distinctive feature systems have poor typological representation of segment inventories
- distinctive feature vectors can be automatically generated for some segment types, however, some "complex" segment types that are undefined by a distinctive feature set must be assigned by hand because feature assignment can be ambiguous, e.g. the features of [p] and [f] do not map straightforwardly to the feature set of [pf]
- with one exception, descriptive universals in phonological systems as stipulated in Hyman 2008 continue to hold on a much larger and broader data set than UPSID_{451}

Lastly, I have fulfilled my aims to:

- create a cross-linguistic data set to undertake phonological typology
- provide novel access to phonological inventories at the feature level
- provide researchers with a tool to undertake phonological typology in ways and with data that were not previously available
- create a typological tool that is extensible and that can interoperate with other sources of linguistic and non-linguistic information
- publish data in open formats
- create avenues for future research

I wrap up the conclusion in my dissertation by discussing how to integrate lexical data into PHOIBLE and I pave avenues of further research in the areas of information theoretic approaches to phonology, measuring complexity of phonological systems and investigating feature-based principles in phonological inventories.

References

Abney, Steven & Steven Bird. 2010. The Human Language Project: Building a Universal Corpus of the World's Languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 88–97.

Atkinson, Quentin D. 2011. Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion From Africa. *Science* 332. 346–359.

Bickel, Balthasar. 2010. Capturing Particulars and Universals in Clause Linkage: A Multivariate Analysis. In Isabelle Bril (ed.), *Clause-Hierarchy and Clause-Linking: The Syntax and Pragmatics Interface*, 51–101. Amsterdam: Benjamins.

Bird, Steven & Gary F. Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. *Language* 79(3). 557–582. `http://www.language-archives.org/documents/portability.pdf`.

Blevins, Juliette. 2009. Another Universal Bites the Dust: Northwest Mekeo Lacks Coronal Phonemes. *Oceanic Linguistics* 48(1). 264–273.

Butcher, Andrew & Marija Tabain. 2004. On the Back of the Tongue: Dorsal Sounds in Australian Languages. *Phonetica* 61. 22–52.

Central Intelligence Agency. 2010. The World Factbook. Tech. Rep. Central Intelligence Agency.

Chanard, C. 2006. Systèmes Alphabétiques Des Langues Africaines. Online: `http://sumale.vjf.cnrs.fr/phono/`.

Chomsky, Noam. 1964. *Current Issues in Linguistic Theory*. The Hague: Mouton.

Chomsky, Noam & Morris Halle. 1965. Some Controversial Questions in Phonological Theory. *Journal of Linguistics* 1. 97–138.

Chomsky, Noam & Morris Halle. 1968. *The Sound Pattern of English*. New York, NY: Harper & Row.

Clements, G. N. 2009. The Role of Features in Phonological Inventories. In Eric Raimy & Charles E. Cairns (eds.), *Contemporary Views on Architecture and Representations in Phonology*, 19–68. MIT Press.

Crothers, John. 1978. Typology and Universals of Vowel Systems in Phonology. In Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik (eds.), *Universals of Human Language Volume 2: Phonology*, 93–152. Stanford University Press.

Crothers, John H., James P. Lorentz, Donald A. Sherman & Marilyn M. Vihman. 1979. Handbook of Phonological Data From a Sample of the World's Languages: A Report of the Stanford Phonology Archive.

Cysouw, Michael, Dan Dediu & Steven Moran. 2012. Still No Evidence for an Ancient Language Expansion From Africa. *Science* 335. 657–b.

Ember, Carol R. & Melvin Ember. 2007. Climate, Econiche, and Sexuality: Influences on Sonority in Language. *American Anthropologist* 109(1). 180–185.

Fought, John G., Robert L. Munroe, Carmen R. Fought & Erin M. Good. 2004. Sonority and Climate in a World Sample of Languages: Findings and Prospects. *Cross-Cultural Research* 38. 27–51.

Halle, Morris. 1962. Phonology in Generative Grammar. *Word* 18(1/2). 54–72.

Hartell, Rhonda L. (ed.). 1993. *Alphabets des langues africaines*. UNESCO and Société Internationale de Linguistique.

Haspelmath, Martin. 2007. Pre-established Categories Don't Exist: Consequences for Language Description and Typology. *Linguistic Typology* 11. 119–132.

Haspelmath, Martin. 2010. Comparative Concepts and Descriptive Categories in Crosslinguistic Studies. *Language* 86(3). 663–687.

Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie. 2008. The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. Available online at `http://wals.info/`.

Haudricourt, André G. 1961. Richesse en phonèmes et richesse en locuteurs. *L'Homme* 1. 5–10.

Hay, Jennifer & Laurie Bauer. 2007. Phoneme Inventory Size and Population Size. *Language* 83. 388–400.

Hayes, Bruce. 2009. *Introductory Phonology*. Blackwell.

Hyman, Larry M. 2008. Universals in Phonology. *The Linguistic Review* 25. 83–137.

Inmon, William H. 1992. *Building the Data Warehouse*. John Wiley & Sons, Inc.

Jakobson, Roman. 1949. On the Identification of Phonemic Entities. *Travaux du Cercle Linguistique de Copenhague* 5. 205–213.

Jakobson, Roman, Gunnar Fant & Morris Halle. 1952. *Preliminaries to Speech Analysis*. MIT Press.

Jakobson, Roman & Morris Halle. 1956. *Fundamentals of Language*. Mouton, The Hague.

Janssen, Dirk, Balthasar Bickel & Fernando Zúñiga. 2006. Randomization Tests in Language Typology. *Linguistic Typology* 10. 419–440.

Johnson, Lawrence. 1976. A Rate of Change Index for Language. *Language in Society* 5(2). 165–172.

Kimball, Ralph. 1996. *The Data Warehouse Toolkit*. John Wiley & Sons, Inc.

Kochetov, Alexei, Sam Al Khatib & Loredana Andreea Kosa. 2008. Areal-typological Constraints on Consonant Place Harmony Systems. In *Paper Presented at 2008 Annual Meeting of the Linguistic Society of America*, .

Lazard, Gilbert. 2006. *La quête des invariants interlangues: la Linguistique est-elle une science?* Paris: Champion.

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the World, Sixteenth Edition*. Summer Institute of Linguistics 16th edn.

Lindblom, Björn & Ian Maddieson. 1988. Phonetic Universals in Consonant Systems. In L. M. Hyman & C. N. Li (eds.), *Language, Speech and Mind: Studies in Honour of Victoria A. Fromkin*, London: Routledge.

LINGUIST List. 2009. Multitree: A Digital Library of Language Relationships. Institute for Language Information and Technology (LINGUIST List), Eastern Michigan University. Ypsilanti, MI. Online: `http://multitree.org/`.

Maddieson, Ian. 1984. *Pattern of Sounds*. Cambridge, UK: Cambridge University Press.

Maddieson, Ian. 1986. The Size and Structure of Phonological Inventories: Analysis of UPSID. In John J. Ohala & Jeri J. Jaeger (eds.), *Experimental Phonology*, Orlando: Academic Press.

Maddieson, Ian. 1991. Testing the Universality of Phonological Generalizations With a Phonetically Specified Segment Database: Results and Limitations. *Phonetica* 48. 193–206.

Maddieson, Ian. 2008a. Consonant Inventories. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library.

Maddieson, Ian. 2008b. Tone. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library.

Maddieson, Ian. 2008c. Vowel Quality Inventories. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library.

Maddieson, Ian & Kristin Precoda. 1990. Updating UPSID. In *UCLA Working Papers in Phonetics*, vol. 74, 104–111. Department of Linguistics, UCLA.

Mithen, Steven. 2003. *After the Ice*. London, UK: Orion Books.

Mukherjee, Animesh, Monojit Choudhury, Anupam Basu & Niloy Ganguly. 2008. Modeling the Co-occurrence Principles of the Consonant Inventories: A Complex Network Approach. *International Journal of Modern Physics C (IJMPC)* 18. 281–295.

Munroe, R. L., R. H. Munroe & S. Winters. 1996. Cross-cultural Correlates of the Consonant-vowel (CV) Syllable. *Cross-Cultural Research* 30. 60–83.

Munroe, Robert L., John G. Fought & Ronald K. S. Macaulay. 2009. Warm Climates and Sonority Classes: Not Simply More Vowels and Fewer Consonants. *Cross-Cultural Research* 43(2). 123–133.

Munroe, Robert L. & Megan Silander. 1999. Climate and the Consonant-Vowel (CV) Syllable: A Replication Within Language Families. *Cross-Cultural Research* 33. 43–62.

Nettle, Daniel. 1999. Is the Rate of Linguistic Change Constant? *Lingua* 108. 119–136.

Newmeyer, Frederick J. 2010. On Comparative Concepts and Descriptive Categories: A Reply to Haspelmath. *Language* 86(3). 688–695.

Pericliev, Vladimir & Raúl E. Valdés-Pérez. 2002. Differentiating 451 Languages in Terms of Their Segment Inven-

tories. *Studia Linguistica* 56(1). 1–27.

Rice, Karen & Peter Avery. 1993. Segmental Complexity and the Structure of Inventories. In *Toronto Working Papers in Linguistics*, vol. 12, University of Toronto.

Rosenfelder, Ingrid, Joe Fruehwald, Keelan Evanini & Jiahong Yuan. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite. Online: `http://fave.ling.upenn.edu`.

Sherman, Donald. 1975. Stop and Fricative Systems: A Discussion of Paradigmatic Gaps and the Question of Language Sampling. In *Working Papers on Language Universals*, vol. 17, 1–31. Stanford University.

Trubetzkoy, Nikolai. 1939. *Grundzüge der Phonologie*. Travaux du cercle linguistique de Prague 7. `http://www.univie.ac.at/Hausa/SprawiODL/Trubetzkoy.html`.

Trudgill, Peter. 1997. Typology and Sociolinguistics: Linguistic Structure, Social Structure and Explanatory Comparative Dialectology. *Folia Linguistica* 31(3–4). 349–360.

Trudgill, Peter. 2002. Linguistic and Social Typology. In J. K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change*, Oxford, UK: Blackwell Publishers.

Wichmann, Søren & Eric W. Holman. 2009. Population Size and Rates of Language Change. *Human Biology* 81. 259–274.