Université Paris Cité

A Statistical Analysis of Vowel Inventories of World Languages

Guillaume Wisniewski
guillaume.wisniewski@u-paris.fr

September 2025

1 Introduction

The aim of this practical session is to demonstrate how data-analysis methods and linguistic databases can be used to compare the characteristics of languages and to discover linguistic knowledge. This is a highly speculative lab : I have no idea what the answers to most of the questions are (indeed, most of them are open-ended). The intention is rather to provide you with a guide to a set of experiments in the hope that (i) you will discover plenty of exciting things (ideally while enjoying yourselves) and (ii) you will acquire useful knowledge and skills — in this case, in data analysis.

One of the aims of this practical session is to examine two well-known proposed linguistic universals relating to vowel systems. First, that virtually all languages possess the basic vowel triangle `[i]`, `[a]`, `[u]` (or close equivalents). Second, that if a language has highly marked vowels (for example nasal, long, or front rounded vowels), it almost always also has the corresponding simpler vowels. These claims are frequently cited in typological and phonetic studies (e.g. (Maddieson 1984 ; Crothers 1978 ; Lindblom et Maddieson 1988)). In this lab, we will test whether these are indeed universals by drawing on `PHOIBLE`, an open, cross-linguistic database of phonological inventories covering over 2,000 languages (Moran et McCloy 2019). `PHOIBLE` aggregates data from multiple published sources and encodes segmental inventories in a standardised format, making it a valuable resource for large-scale comparisons of vowel systems across languages.

2 Getting Started with `PHOIBLE`

1. Determine how many different languages are included in `PHOIBLE`, the average number of vowels per language, and the distribution of vowel inventory sizes.

2. Explain why it is interesting to examine these statistics.
3. Identify outliers using the Interquartile Range (IQR) method. Briefly explain the method and comment on your findings.

3 Testing Linguistic Universals

The basic idea of this part is to use the data to test whether the proposed universals are actually observed across the sample of languages in the `PHOIBLE` database.

4. Identify the three most frequent vowels (those occurring in the largest number of languages).
5. Compute the proportion of languages that contain specific vowels the basic vowel triangle `[i]`, `[a]`, `[u]` and discuss whether this supports the universals.

We will now investigate whether two phonemes tend to occur together in the same languages more often than would be expected by chance. For each pair (A, B) of phonemes :

6. Construct a 2×2 contingency table indicating, for each language in the dataset :
   — whether it has phoneme A,
   — whether it has phoneme B.
   The four cells of the table will then correspond to : (A present & B present), (A present & B absent), (A absent & B present), (A absent & B absent).
7. Compute a χ^2 test of independence on this table. Does the result suggest that the two phonemes occur independently, or that there is a statistically significant association ?
8. Compute the *mutual information* between the two phonemes, defined as

$$I(A;B) = \sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a,b) \log \frac{p(a,b)}{p(a)\,p(b)}, \tag{1}$$

   where $p(a,b)$ is the proportion of languages with the configuration (a,b), and $p(a)$ and $p(b)$ are the marginal probabilities of each phoneme. Interpret the value : is it close to zero (independence), positive (positive association), or negative (mutual exclusion) ?
9. Compare the results of the two methods (χ^2 and mutual information). Do they lead to similar conclusions about the relationship between the two phonemes ?
10. What can you conclude ?

4 Discovering Regularities in the Corpus

The idea now is to see whether it is possible to identify other regularities from the data contained in the corpus. To do this, we will begin by looking at whether there are any *frequent itemsets*.

A frequent itemset is a set of items (in our case, features or phonemes) that occur together in the data more often than a given minimum threshold. In the context of language inventories, an item could be the presence of a particular phoneme, and an itemset would be a combination of phonemes appearing in the same language. For example, the pair {`[i]`, `[u]`} would form an itemset ; if this pair appears in many languages, it can be considered a frequent itemset.

Frequent itemsets are evaluated using support, which measures how often the itemset appears relative to the total number of observations. Formally, the support of an itemset X is defined as :

$$\text{support}(X) = \frac{\text{number of records (languages) containing } X}{\text{total number of records (languages)}} \tag{2}$$

Only itemsets whose support exceeds a chosen *minimum support threshold* (e.g. 10% of all languages) are considered frequent. This approach is commonly used in association-rule mining (such as the `Apriori` or `FP-Growth` algorithms) to uncover systematic co-occurrence patterns in large datasets.

Once frequent itemsets have been identified, it is often useful to go further and examine **association rules** between items. Two key measures used for this purpose are *confidence* and *lift*.

Confidence measures how often items in Y appear in transactions (or here, in languages) that already contain X. It reflects the conditional probability of Y given X. Formally, for a rule $X \to Y$:

$$\text{confidence}(X \to Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} = P(Y|X)$$

A high confidence value means that when X occurs, Y also occurs in a large proportion of cases.

Lift compares the observed co-occurrence of X and Y with what would be expected if they were statistically independent. It tells us whether the presence of X increases (lift > 1), decreases (lift < 1), or has no effect on (lift = 1) the likelihood of Y. Formally :

$$\text{lift}(X \to Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X) \times \text{support}(Y)}$$

Lift thus normalises confidence by the baseline frequency of Y.

In the context of phoneme inventories, high confidence for a rule such as {/i/} → {/u/} would mean that most languages containing /i/ also contain /u/. A lift greater than 1 would indicate that /i/ and /u/ co-occur more often than expected by chance.

Using the `PHOIBLE` dataset :

11. For each possible itemset of size 1 (single phonemes), compute its *support* (proportion of languages containing that phoneme).
12. For all itemsets of size 2 (pairs of phonemes), compute their *support* in the same way.
13. Define as **frequent itemsets** those whose support exceeds the *minimum support threshold* chosen for this practical (e.g. 10% of all languages).
14. Identify the frequent itemsets of size 2 and comment on any regularities you observe.

15. For at least one of the frequent itemsets you identified, compute the *confidence* and *lift* for the association rule $X \rightarrow Y$ where X and Y are individual phonemes within that itemset.
16. Interpret your results : Which phoneme combinations are most typical across languages ? Do the confidence and lift values indicate stronger or weaker associations than expected under independence ?

To go beyond manually calculating support for pairs of phonemes, you can use an existing implementation of the **Apriori algorithm**, which efficiently discovers frequent itemsets of any size. One convenient implementation is provided in the Python library `mlxtend` (Machine Learning Extensions). Its `apriori` function takes a binary matrix (rows = languages, columns = presence/absence of phonemes) and returns all itemsets above a specified minimum support threshold. The same library also includes a `association_rules` function for computing confidence and lift automatically.

17. Prepare your dataset as a binary matrix indicating, for each language, whether each phoneme is present (1) or absent (0).
18. Use the `apriori` function to identify frequent itemsets of size 2 and, if you wish, larger itemsets by adjusting the minimum support threshold.
19. Export the resulting itemsets together with their support values.
20. Using the `association_rules` function, compute the confidence and lift for at least two association rules of your choice and comment on your findings.

Références

Crothers, John (1978). *Typology and Universals of Vowel Systems in Human Languages*. Rapp. tech. Stanford, CA : Stanford University.

Lindblom, Björn et Ian Maddieson (1988). "Phonetic universals in consonant systems". In : *Language, Speech and Mind : Studies in Honour of Victoria A. Fromkin*. Sous la dir. de Larry M. Hyman et Charles N. Li. London : Routledge, p. 62-78.

Maddieson, Ian (1984). *Patterns of Sounds*. Cambridge : Cambridge University Press.

Moran, Steven et Daniel McCloy (2019). *PHOIBLE 2.0*. Available online at https://phoible.org. Max Planck Institute for the Science of Human History.