PhonMatrix: Visualizing co-occurrence constraints of sounds

Thomas Mayer Research Unit Quantitative Language Comparison Philipps University of Marburg thomas.mayer@uni-marburg.de

Abstract

This paper describes the online tool Phon-Matrix, which analyzes a word list with respect to the co-occurrence of sounds in a specified context within a word. The cooccurrence counts from the user-specified context are statistically analyzed according to a number of association measures that can be selected by the user. The statistical values then serve as the input for a matrix visualization where rows and columns represent the relevant sounds under investigation and the matrix cells indicate whether the respective ordered pair of sounds occurs more or less frequently than expected. The usefulness of the tool is demonstrated with three case studies that deal with vowel harmony and similar place avoidance patterns.

1 Introduction

In this paper, we introduce the *PhonMatrix*¹ tool, which is designed to visualize co-occurrence constraints of sounds within words given a reasonably sized word list of the language. It is a web-based implementation of the visualization method proposed in (Mayer et al., 2010a), including some further development such as an interactive component and a range of association measures and sorting methods to choose from. The original motivation for this tool is to give linguists the opportunity to upload their own word lists in order to visually explore co-occurrence constraints in languages. The basic idea behind the visual component of the tool is to provide for a first, at-a-glance mode of analysis which can be used to generate hypotheses about the data by simply looking at the visualization matrices.

Christian Rohrdantz Data Analysis and Visualization Group University of Konstanz christian.rohrdantz @uni-konstanz.de

Phonotactic constraints in languages abound. One of the most well-known and wide-spread constraints is commonly referred to as vowel harmony (van der Hulst and van de Weijer, 1995). In vowel harmony languages, vowels are separated into groups where vowels of the same group tend to co-occur within words, while vowels from different groups rarely co-occur. Likewise, in some languages there are patterns of consonant harmony (Hansson, 2010) that show a similar behavior with respect to consonants. Less common are cases of "synharmonism" (Trubetzkoy, 1967, p. 251) where both vowels and consonants form such groups and words usually only contain sounds from the same group (e.g., only front vowels and palatalized consonants). Whereas vowel harmony patterns are easily detectable in many harmonic languages due to the harmonic alternants in affixes, other co-occurrence constraints are less obvious. This is especially true for disharmonic patterns, the most famous of which is the principle of Similar Place Avoidance (SPA) in Semitic consonantal roots (Greenberg, 1950). Recent studies have shown that this principle is not only active in Semitic languages, where it was already known by grammarians in the Middle Ages, but is a more widespread tendency among the languages of the world (Pozdniakov and Segerer, 2007; Mayer et al., 2010b). One of the reasons why statistical constraints like SPA are more difficult to detect is the fact that they exhibit many frequent counterexamples and are therefore less easily spotted as a general (albeit statistical) tendency.

In our view, there are many more phonotactic constraints that wait to be discovered by linguists. With the availability of language data in electronic format such tendencies can be automatically processed and presented to the user in a form that allows for an easy exploration of the results in a short period of time. Thus a larger number of phonotactic contexts can be explored in order to

¹http://paralleltext.info/phonmatrix/

find potential patterns in the data. The *PhonMatrix* tool is part of an ongoing effort to integrate methods and techniques from the field of visual analytics (Thomas and Cook, 2005) into linguistic research. The present tool will be gradually augmented with further functionalities in order to enhance its usefulness.

2 Related work

A related tool that quantifies the co-occurrence of sounds in a given corpus is the Vowel Harmony Calculator (Harrison et al., 2004). The major difference between PhonMatrix and the Vowel Harmony Calculator is that the latter is restricted to the context of vowel harmony and requires the user to input the harmony classes beforehand whereas PhonMatrix is designed to detect these classes by making potential harmonic patterns more easily perceptible to the user. The Vowel Harmony Calculator quantifies the notion of vowel harmony for the input corpus by giving the percentage of harmonic words and the harmony threshold. The harmony threshold is the percentage of words that would be expected to be harmonic purely by chance. The output of the Vowel Harmony Calculator consists of a list of values (number of polysyllabic words, harmony threshold, percentage of harmonic words, harmony index, among other things) but does not give any information about the harmonic strength of individual vowel pairs. In short, the Vowel Harmony Calculator is a way to quantify the notion of harmony given the harmony classes of the language whereas Phon-Matrix is intended to help detect such patterns.

3 System overview

PhonMatrix is a web-based visualization tool that statistically analyzes sound co-occurrences within words and displays the result in a symmetric sound matrix. The statistical components are written in Python whereas the visualization part is in Javascript, using the D3 library (Bostock et al., 2011). Before discussing the individual steps of the system in more detail we will give a brief overview of the overall processing pipeline (see Figure 1).

In the first step, the user has to upload the text file containing the word list that serves as the input to the analysis process. Text files have to be encoded in UTF-8 and list only one word per line. For a meaningful analysis the words should be given in some phonemic transcription (e.g., using IPA).²

After the file has been uploaded to the server all symbols in the word list are analyzed according to their unigram and bigram frequencies. These frequencies are used to infer an automatic distinction between vowels, consonants and infrequent symbols. Infrequent symbols are considered to be noise in the data and can be ignored for further processing. A distinction between vowels and consonants is automatically inferred from the word list by means of Sukhotin's algorithm (Sukhotin, 1962). The results of Sukhotin's algorithm are presented to the user together with the frequency counts of the individual symbols in the word list.

In the third step, the user can make changes to the automatic classification of symbols into vowels and consonants and exclude infrequent symbols from further consideration. The subsequent calculations of co-occurrence values are mostly based on the distinction of input symbols into consonants (C) and vowels (V). Users can choose among a number of options that define the context for the co-occurrence calculations.³ Two of those options will be discussed in more detail in this paper (vowel co-occurrences in VCV and CC sequences). Depending on the user's choice, the co-occurrences in the selected context are calculated and analyzed with respect to a number of statistical association measures from which the user can choose one for the visualization.

In the last step, the results of the statistical analysis of the co-occurrence counts are displayed in a quadratic matrix of sounds. The rows and columns of the matrix represent the individual sounds that are relevant for the selected context (e.g., vowels in the context of VCV sequences). The rows thereby stand for the first members of the relevant sound pairs, whereas the columns contain the second members. Each cell of the matrix then shows the result for the pair of sounds in the respective row and column.

The final result is a visualization of the cooccurrence matrix with rows and columns sorted according to the similarity of the sound vectors and statistical values represented as colors in the matrix cells. The visualization features a number

²For more information on the minimum amount of data necessary see (Mayer et al., 2010a).

³It is also possible for users to define their own contexts with regular expressions.



Figure 1: The processing pipeline of the PhonMatrix visualization tool.

of interactive components that facilitate the detection of potential patterns in the results by the user.

4 **PhonMatrix components**

PhonMatrix consists of three main components: preprocessing (including vowel-consonant distinction), statistical analysis of co-occurrence counts and visualization. In what follows, we will describe each component in more detail, with special emphasis on the visualization component.

4.1 Vowel-consonant distinction

Most of the co-occurrence restrictions that might be of interest make reference to a distinction between vowels and consonants. Since a manual classification of all sounds in the input into vowels and consonants is a tedious task (especially with a larger number of symbols), the first component deals with an automatic inference of such a distinction. Many methods have been discussed in the literature on how to discriminate vowels from consonants on the basis of their distribution in texts. Many of them involve many lines of code and are computationally demanding. Yet there is a very simple and fast algorithm that yields reasonably good results (Sukhotin, 1962; Guy, 1991).

The basic idea of Sukhotin's algorithm is that vowels and consonants have the tendency not to occur in groups within words but to alternate. Based on the additional assumption that the most frequent symbol in the text is a vowel, the algorithm iteratively selects the symbol which occurs most frequently adjacent to a vowel and determines it to be a consonant. The algorithm stops if no more consonants can be selected because no co-occurrence counts with any remaining vowel are positive. Although the algorithm is quite old and very simple, it gives reasonably good results (Goldsmith and Xanthos, 2009; Guy, 1991; Sassoon, 1992). PhonMatrix makes use of Sukhotin's algorithm as a preprocessing step to give a first guess of the class for each symbol, which the user can then modify if it turns out to be wrong. It mainly serves to speed up the classification step.

4.2 Co-occurrence statistics

With the distinction of symbols into vowels and consonants at hand, the user can then select a relevant context for the co-occurrence counts. The relevant context can be chosen from a list of predefined options. Here we will illustrate the statistical analysis with the context of VCV sequences to investigate vowel harmony in Turkish. The input consists of 20,968 orthographic words from the Turkish New Testament.⁴ The tool automatically extracts all VCV sequences in the words and counts the co-occurrences of sounds in these sequences. The counts are then summarized in a quadratic contingency table and can be used for further statistical analyses.

In our experiments, two measures turned out to be especially useful for the detection of potential patterns: the probability and ϕ values. The ϕ value is a normalized χ^2 measure which allows for an easier mapping of values to the color scale because it is always between -1 and $1.^5$ The ϕ values for the vowels in the Turkish text are shown in Table 1. Apart from probability and ϕ values, the user can also choose among a number of other association measures such as pointwise mutual information, likelihood ratios or t-scores (Manning and Schütze, 1999).

4.3 Visualization component

The input to the visualization component is a matrix of association measures for each sound pair in the relevant context. Two additional steps have to be performed in order to arrive at the final matrix visualization: 1) the rows and columns of the matrix have to be sorted in a meaningful way; 2) the association measures have to be mapped to visual variables. For the matrix arrangement, we decided to have the same order of symbols for the rows and columns. The order of symbols is determined by a clustering of the

⁴Turkish orthography represents the modern pronunciation with a high degree of accuracy.

⁵Apart from this, ϕ makes good use of the off-diagonal cells in the contingency tables (Church and Gale, 1991).

	а	e	i	0	u	ö	ü	1
а	0.53699	-0.49730	-0.54579	-0.30421	-0.38117	-0.03895	-0.36874	0.65791
e	-0.48371	0.54763	0.64548	-0.28216	-0.37907	-0.05792	-0.32882	-0.53454
i	-0.40334	0.37477	0.59682	0.30227	-0.33970	0.09038	-0.30307	-0.49651
0	0.20048	-0.28306	-0.31395	-0.14114	0.65493	-0.05532	-0.20696	-0.33238
u	0.28855	-0.34937	-0.38283	0.17629	0.73451	0.10011	-0.22066	-0.39304
ö	-0.28879	0.32352	-0.29843	-0.16465	-0.21329	-0.04885	0.65373	-0.29354
ü	-0.31709	0.33094	-0.34774	0.14995	-0.24351	-0.05829	0.75780	-0.35024
1	0.30302	-0.40711	-0.46423	0.32671	-0.33210	-0.07607	-0.28459	0.58548

Table 1: ϕ values of VCV sequences in Turkish.

symbols based on the similarity of their row values. The clustering is performed with the Python scipy.cluster.hierarchy package from the SciPy library. As a default setting Ward's algorithm (Ward, 1963) is used but other clustering algorithms can also be easily integrated.

Whereas the preprocessing steps and the datadriven sorting of rows and columns have been written in Python, the actual visualization of the results in the browser is implemented in Javascript using the D3 library (Bostock et al., 2011). The association measures and the order of the symbols are referenced as Javascript variables in the visualization document. The data is then automatically bound to DOM elements of the HTML document through the D3 data operator. The mapping from association measures to color values is made with the linear scale method from the d3.scale package. Scale methods map from an input domain to an output range. The input domain for the ϕ values is the interval [-1;1], while the output range can be given as a color scale ranging from one color to the other. For the ϕ values we decided to use two unipolar scales, one from -1 to 0 (red) and the other from 0 to +1 (blue). In order to reserve a larger color range for the densely populated area of low values we did not linearly map the numerical association measures but used the square roots of the numerical values as the input for the scale function. Additionally, the sign of the ϕ value, which shows whether the co-occurrence of a certain symbol pair occurs more (+) or less (-) frequently than expected, is displayed in the matrix cell.⁶ The result of the matrix visualization for the ϕ values of the vowels in Turkish VCV sequences is shown in Section 5.1.

The matrix visualization also features some interaction to explore the results in more detail. On mouse-over, the respective matrix cell shows the actual values that serve as the input for the data mapping process. Additionally, the row and column labels are highlighted in order to show more clearly which pair of symbols is currently selected (see Figure 2). The size of the matrix can also be adjusted to the user's needs with the help of a slider above the matrix. Next to the slider is a dropdown menu from which users can choose the association measure that they want to be displayed in the visualization.

5 Case studies

After the description of the PhonMatrix system we will illustrate the usefulness of the visualization of co-occurrence patterns in sounds with three case studies. They are presented as a proof of concept that the visualization component allows for an ata-glance exploration of potential patterns. The visualization part is thereby not considered to be a replacement of more detailed linguistic investigations but rather serves as a way to explore a multitude of different contexts and data in a comparatively short period of time. After a suspicious pattern has been detected it is indispensable to look at the actual data to see whether the visualization result is an artifact of the method or data at hand or whether the detected pattern is an interesting phonotactic feature of the language under consideration.

5.1 Turkish vowel harmony

The first case study shows the results of the VCV sequences in Turkish described above. For this purpose the vowels a, e, i, o, u, \ddot{o} , \ddot{u} , ι are selected as the relevant sounds that are to be compared in

⁶The algebraic sign is displayed in white and therefore stands out more clearly with higher absolute ϕ values.

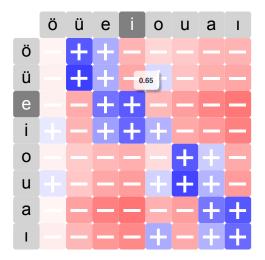


Figure 2: The visualization of the ϕ values of VCV sequences in the Turkish text.

the visualization. Figure 2 shows the results for the ϕ values that have been computed from the co-occurrence counts of the symbols in VCV sequences. The arrangement of the symbols in the matrix rows and columns already show a distinction between front (the first four vowels) and back (the last four vowels) vowels, reflecting the palatal harmony in Turkish. This distinction can best be seen when looking at the e- and a-columns where the top four vowels all have positive ϕ values for e and negative ϕ values for a, whereas the bottom four vowels show the opposite behavior. On closer inspection, the labial harmony for high vowels can also be seen in the matrix visualization. From top to bottom there are always pairs of vowels that take the same harmonic vowel, starting with (ö, \ddot{u}) taking \ddot{u} and followed by (e, i) taking i, (o, u)taking u and finally (a, i) taking i. The usefulness of the visualization component to detect such patterns can best be seen when comparing Figure 2 with Table 1, which contains the same information.

5.2 Finnish vowel harmony

The second case study shows that the harmonic patterns can also be detected in orthographic words of the Finnish Bible text. Finnish differs from Turkish in having only one type of harmony (palatal harmony) and neutral vowels, i.e., vowels that do not (directly) participate in the harmony process. As a different underlying association measure for the visualization consider the probability values in Figure 3. For probability values

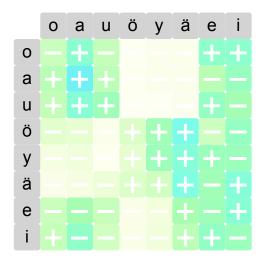


Figure 3: The visualization of the probabilities of VCV sequences in the Finnish text.

we have chosen a bipolar color scale ranging from white (for 0) over green (for 0.5) to blue (for 1). The probability matrix clearly shows the relevant blocks of vowels that mark the harmony groups.⁷ The clustering algorithm separates the back vowels (first three vowels *o*, *a*, *u*) from the front vowels (vowels four to six, \ddot{o} , *y*, \ddot{a}) and the neutral vowels (*e*, *i*). The blocks along the main diagonal of the matrix show the harmonic pattern among the harmony groups, whereas the neutral vowels do not display any regular behavior.

5.3 Maltese verbal roots

PhonMatrix is not only useful to find vowel harmony patterns. The third case study shows that other co-occurrence constraints such as SPA can also be detected. To illustrate this, we show the visualization of CC patterns in a comprehensive list of Maltese verbal roots (Spagnol, 2011). The consonant matrix in Figure 4 shows two clusters, with one cluster (the first twelve consonants in the top row) containing labial and dorsal and the other cluster (the last eleven consonants) comprising only coronal consonants.⁸ The visualization also reveals that, unlike in vowel harmony, consonants from the same cluster do not occur next to each other in the CC sequences, as shown by the red blocks in the top left and bottom right. This is exactly what SPA would predict.

⁷The +/- signs in the matrix are taken from the ϕ values. ⁸The consonants are given in their orthographic representation (Borg and Azzopardi-Alexander, 1997, p. 299).

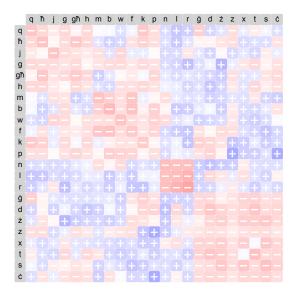


Figure 4: The visualization of the ϕ values of consonant sequences in Maltese verbal roots.

6 Conclusions

In this paper, we have presented *PhonMatrix*, a web-based, interactive visualization tool for investigating co-occurrence restrictions of sounds within words. The case studies of vowel harmony and SPA have shown that interesting patterns in the data can easily be seen only by looking at the matrix visualizations.

Acknowledgments

This work was partially funded by the DFG project "Algorithmic corpus-based approaches to typological comparison." We are grateful to two anonymous reviewers for their valuable comments and suggestions.

References

- Albert Borg and Marie Azzopardi-Alexander. 1997. *Maltese*. Descriptive Grammar Series. London: Routledge.
- Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3: Data-driven documents. *IEEE Transactions on Visualization & Computer Graphics (Proc. InfoVis)*, 17(12):2301–2309.
- Kenneth W. Church and William A. Gale. 1991. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62.
- John Goldsmith and Aris Xanthos. 2009. Learning phonological categories. *Language*, 85(1)(1):4–38.

- Joseph H. Greenberg. 1950. The patterning of root morphemes in Semitic. *Word*, 6:161–182.
- Jacques B. M. Guy. 1991. Vowel identification: an old (but good) algorithm. *Cryptologia*, 15(3):258–262, July.
- Gunnar Ólafur Hansson. 2010. *Consonant Harmony*. Berkeley: University of California Press.
- David Harrison, Emily Thomforde, and Michael O'Keefe. 2004. The vowel harmony calculator. http://www.swarthmore.edu/ SocSci/harmony/public_html/.
- Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Natural Language Processing. Massachusetts Institute of Technology.
- Thomas Mayer, Christian Rohrdantz, Miriam Butt, Frans Plank, and Daniel A Keim. 2010a. Visualizing vowel harmony. *Journal of Linguistic Issues* in Language Technology (LiLT), 4(2):1–33.
- Thomas Mayer, Christian Rohrdantz, Frans Plank, Peter Bak, Miriam Butt, and Daniel A. Keim. 2010b. Consonant co-occurrence in stems across languages: Automatic analysis and visualization of a phonotactic constraint. In *Proceedings of the ACL 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 67–75.
- Konstantin Pozdniakov and Guillaume Segerer. 2007. Similar Place Avoidance: A statistical universal. *Linguistic Typology*, 11(2)(2):307–348.
- George T. Sassoon. 1992. The application of Sukhotin's algorithm to certain Non-English languages. *Cryptologia*, 16(2)(2):165–173.
- Michael Spagnol. 2011. A tale of two morphologies. Verb structure and argument alternations in Maltese. Ph.D. thesis, Germany: University of Konstanz dissertation.
- Boris V. Sukhotin. 1962. Eksperimental'noe vydelenie klassov bukv s pomoşçju evm. Problemy strukturnoj lingvistiki, 234:189–206.
- James J. Thomas and Kristin A. Cook. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr.
- N. S. Trubetzkoy. 1967. *Grundzüge der Phonologie*. Göttingen: Vandenhoeck & Ruprecht. 4. Auflage.
- Harry van der Hulst and Jeroen van de Weijer. 1995. Vowel harmony. In John Goldsmith, editor, *The Handbook of Phonological Theory*, chapter 14, pages 495–534. Basil Blackwell Ltd.
- Joe H. Jr. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(1)(1):236–244.