First publ. in: NLPLING 2010 : 2010 Workshop on NLP and Linguistics, Finding the Common Ground, Proceedings of the Workshop. Stroudsburg, PA : Association for Computational Linguistics (ACL), 2010, pp. 70 - 78

Consonant Co-occurrence in Stems Across Languages: Automatic Analysis and Visualization of a Phonotactic Constraint

Thomas Mayer[1], **Christian Rohrdantz**[2], **Frans Plank**[1],
Peter Bak[2], **Miriam Butt**[1], **Daniel A. Keim**[2]
[1]Department of Linguistics, [2]Department of Computer Science
University of Konstanz, Germany
{thomas.mayer,christian.rohrdantz}@uni-konstanz.de

Abstract

In this paper, we explore the phenomenon of Similar Place Avoidance (SPA), according to which successive consonants within stems sharing the same place of articulation are avoided. This principle has recently been hypothesized as a universal tendency although evidence from only a few languages scattered across the world has been considered. Using methods taken from the field of Visual Analytics, which have demonstrably been shown to help with understanding complex interactions across large data sets, we investigated a large crosslinguistic lexical database (comprising data on more than 4,500 languages) and found that a universal tendency can indeed be maintained.

1 Introduction

Linguistic knowledge has traditionally been acquired by analyzing a manageable set of data, on the basis of which generalizations are posited that can then be tested on an extended set of data from the same language or comparative data from other languages. Tendencies, rather than absolute principles, are difficult to detect under this approach. This is true especially when they are obscured by counterexamples that happen to occur with high frequency, but that may be restricted to just a small minority of the overall pattern. This may prompt a researcher to discard a valid generalization from the outset. In recent years, a plethora of statistical and stochastic methods have therefore been pursued within linguistic research, leading to approaches such as stochastic Optimality Theory (Boersma and Hayes, 2001) or the use of statistics to detect crosslinguistic tendencies (Bickel, in press).

However, although the various statistical methods deal with data which exhibit very complex and often ill-understood interactions, analyses have not to date availed themselves of methodology currently being developed in the field of Visual Analytics, which allows us to use our powerful visual processing ability to understand and evaluate complex data sets (Keim et al., 2008; Thomas and Cook, 2005).

In this paper, we present an interdisciplinary effort whereby linguistically interesting patterns are automatically extracted, analyzed and visually presented so that an at-a-glance evaluation of linguistically significant patterns is made possible. In order to demonstrate that this technique is especially useful with phenomena that do not manifest themselves in absolute principles, but rather in statistical tendencies, we investigated a phenomenon that, on the basis of a comparatively sparse and unrepresentative data set, has recently been claimed to be a universal tendency (Pozdniakov and Segerer, 2007): *Similar Place Avoidance* (SPA). In this paper, we conduct a more representative study of about 4,500 languages. Our results allow an at-a-glance evaluation which shows that SPA indeed seems to be a valid language universal tendency.

Our work on SPA is part of a more widespread effort currently being conducted with respect to visually representing crosslinguistic sound patterns. In Rohrdantz et al. (2010), we already showed that phonological patterns in languages can be automatically extracted and visualized from corpora. Figure 1 displays the vowel harmony patterns that were extracted for Turkish in comparison with the lack of such patterns in a non-harmonic language like Spanish.

The remainder of this article is organized as follows. Section 2 introduces SPA. Section 3 provides an overview of the material that was used. A description of the calculations and statistical analyses is given in Section 4. Section 5 presents the results of the geo-spatial visualizations, partly

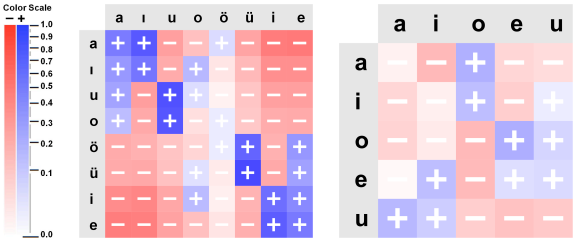

Figure 1: Turkish vowel harmony patterns (left). The matrix visualizaton was generated on the basis of the Turkish Bible text and shows the palatal (front/back) and labial (rounding) harmony blocks. Rows and columns are automatically sorted according to the similarity of vowels. For non-harmonic languages, such as Spanish (right), no such patterns can be detected.

with respect to a WALS map (Haspelmath et al., 2005). In the final section, we consider some implications of our findings and raise some questions for future research.

2 Similar Place Avoidance (SPA)

It has long been noted in studies on Semitic languages, especially Arabic, that there are constraints on the structure of triliteral consonant roots ($\sqrt{CCC}$) with respect to the phonological features of the individual consonants (Greenberg, 1950). The basic observation is that consonants with a similar place of articulation are avoided in non-derived forms. A similar observation has also been made with respect to the Proto-Indo-European (PIE) roots. Among other things, Iverson and Salmons (1992) note that Stop-V-Stop roots were very rare in PIE, representing only 3.5% of a lexicon of more than 2,000 items. Plank (1981:221f) observes that Modern German tends to avoid verbal stems with identical consonants in initial and final positions (allowing for differences in voicing), and that those verbs with identical initial and final consonants which do exist are all morphologically regular. This indicates that they are not basic verbs, but represent a technique of word formation, perhaps derivative of reduplication as especially common in child or child-directed language.[1]

[1]Note that the early speech of children is characterized by the opposite effect of SPA: both consonants and vowels tend to share the same place of articulation (Fikkert and Levelt, 2010), with greater and greater differentiation being achieved in the course of language acquisition. The reasons for this remain to be investigated.

Looking at suprasegmental features, Leben (1973) argued that a similar restriction holds for the co-occurrence of tones in underlying representations. In the framework of Autosegmental Phonology this has become known as the *Obligatory Contour Principle* (OCP), which precludes sequences of identical tones from underlying representations. This principle has since been understood more generally as a prohibition on similar items and has thus also been used in relation with the SPA bias in Semitic radicals.

More recently, the application of SPA with respect to stem-internal consonants has been claimed for other non-Semitic languages as well. Pozdniakov and Segerer (2007) found impressive support for it in their sample of Atlantic and Bantu languages of Niger-Congo and further tested its crosslinguistic validity for some more languages or language groups (Mande, Kwa, Ubangi, Sara-Bongo-Bagirmi, Chadic, Malagasy, Indo-European, Nostratic, Mongolian, Basque, Quechua, Kamilaroi, Port Moresby Pidgin English) with similar results. Table 1 shows their findings across all 31 languages in their sample. It can be seen that the highest negative numbers are in the main diagonal of the matrix, which is exactly what SPA would predict.

	P	T	C	K
P	−15	+11	+5	−5
T	+12	−10	−5	+13
C	+8	−5	−6	+8
K	−3	+8	+5	−15

Table 1: Results in Pozdniakov and Segerer (2007). The numbers indicate the overall sum of cells with negative vs. positive values with regard to successions of places of articulation (see Section 3 for a description of the labels *P, T, C* and *K*) for all languages in their sample. Positive and negative values have been assigned if the observed absolute value was at least 15% above (respectively below) the expected value. Compare their results with the left matrix in Figure 3.

3 Database and methodology

The data that underlies all the subsequent work presented in this paper have been taken from the Automated Similarity Judgment Program (ASJP; Wichmann et al., 2010), which aims at achiev-

ing a computerized lexicostatistical analysis of the world's languages. To this end, Wichmann and his collaborators have collected Swadesh list items for over 4,500 languages. The so-called Swadesh list was developed by Morris Swadesh in the 1940–50s with the aim of having a basic set of vocabulary items which are culturally neutral and which one would expect to be stable over time. The original idea of a Swadesh list was to be able to compare and test languages with respect to genealogical relations.

The Swadesh items in the Wichmann et al. database are transcribed in the ASJP orthography, which uses standard ASCII characters to encode the sounds of the world's languages, but does merge some of the distinctions made by the IPA. Furthermore, stress, tone and vowel length are not recorded in the database. However, for the purpose of our investigation the transcription is suitable because place of articulation is sufficiently distinguished.

We decided to experiment with two different approaches for dividing up the place of articulation features. One approach (PTCK) is based on the arrangement in Pozdniakov and Segerer (2007) and distinguishes four places of articulation for labial (*P*), dental (and alveolar) (*T*), (alveo-)palatal (*C*) and velar (*K*) consonants. A second grouping (LCD) only distinguishes three places of articulation: labial (*L*), coronal (*C*) and dorsal (*D*).[2] According to this classification the consonants of all the items in the database can be assigned to one of these symbols, as shown in Table 2.

LCD	PTCK	ASJP	IPA
L	*P*	p, b, m, f, v, w	p, ɸ, b, β, m, f, v, w
C	*T*	8, 4, t, d, s, z, c, n, S, Z	θ, ð, n̪, t, d, s, z, ts, dz, n, ʃ, ʒ
	C	C, j, T, l, L, r, y	ʧ, ʤ, c, ɟ, l, ʎ, ɭ, ʎ, r, ɾ, j
D	*K*	5, k, g, x, N, q, G, X, 7, h	ɲ, k, g, x, ɣ, ŋ, q, ɣ, χ, ʁ, ħ, ʕ, ʔ, h, ɦ,

Table 2: Assignment of consonants to symbols. All varieties of "click"-sounds have been ignored.

[2]Radical and laryngeal, which are commonly employed in the phonological literature as yet another place distinction, are subsumed under dorsal.

Experiments with using the four-way distinction vs. the three-way distinction showed that *T* and *C* in the four-way grouping behave very similarly with respect to the transitions to other places of articulation (see Section 4.2). We therefore decided to use the three-way distinction for the bulk of our calculations and visualizations and only sporadically resort to the four-way grouping when a more fine-grained distinction is needed.

Furthermore, we decided to only include those cases where the first and second consonants are preceded (or followed, respectively) by another vowel or a word boundary and are therefore not part of a consonant cluster. We mainly did this in order to minimize the noise caused by consonants of inflectional markers that tend to assimilate in such clusters.

In the literature on root morphemes in Semitic, it has been noted that the consonants within triliteral radicals behave differently with respect to OCP. Greenberg (1950:162) remarks that while the first and second consonants are usually not identical, the same does not hold for the second and third consonants, which frequently constitute the well-known geminate subtype of Semitic verbs. However, for our work we understand OCP as it was later formulated within the framework of autosegmental phonology (Leben, 1973; McCarthy, 1986; Goldsmith, 1976) in that adjacent identical elements (here in the sense of identical with respect to place of articulation) are prohibited, under the assumption that consonants are adjacent to each other (on the C tier) even when they are separated by vowels in the linear sequence of phonemes within the word.

For the purposes of our experiment, we considered the relevant context for adjacency to be one where consonants are separated by exactly one vowel.[3] Note that since the basis for our calculations were not stems in the language but the citation forms that are used in the Swadesh lists, we also get noise from inflectional markers that are attached to these forms and might have the same place of articulation irrespective of the stem to which they attach.[4]

Finally, there are several shortcomings of the

[3]Since vowel length is not marked in the ASJP database, long vowels are also included.

[4]Assimilation processes are far more frequent than dissimilation processes in this context so that it is more likely that the same place of articulation features are to be expected when an inflectional marker is present.

material in the database with respect to our investigation which must be kept in mind. OCP/SPA has been claimed to apply with respect to underlying or non-derived representations. Previous work has been done on the basis of stem (or root) lists. Depending on the language, Swadesh list items are not always stems, but whole words in their citation forms. For instance, while both English and German use the infinitive as the citation form for verbal stems, in English the infinitive is identical to the stem whereas in German it is marked with the suffix *-en*. In other languages, verbs can also be cited by inflected forms other than the infinitive (e.g., the 3rd person singular perfective in Arabic, or the first person singular indicative present in Latin). The same holds for nouns or other word classes that are included in the Swadesh list. Another problematic aspect is the fact that it also contains items (such as personal pronouns) that are not lexical in the strict sense of the meaning and are realized as bound forms in many languages.

Apart from that, the number of items for each language in the ASJP database varied greatly from only a few to one hundred. Moreover, the number of CVC sequences within the items differed greatly from one language to another, depending on the phonotactic properties of the languages. Previous statistical studies have relied on a much larger number of stems and consonant sequences. Pozdniakov and Segerer's (2007) statistics, for example, were calculated on the basis of 495 to 17,944 CVC successions for the languages in their sample.[5] In contrast, our statistics are based on much fewer CVC successions, ranging from 21 to 246 per language. Nevertheless, our results actually correspond to the main findings of their study so that we think that the data are good enough for our purposes.

4 Automated statistical analysis

4.1 Methodology

In a first step, for each language in the sample an elementary statistical processing is performed. Thereby, all successions of places of articulation occurring in the Swadesh list items are identified and counted. To do so, we define a succession of places of articulation as a binary sequence of consonants (C-C). These consonants have to appear within a word and have to be separated by exactly one vowel (V). Before and after the succession either word boundaries (#) or vowels have to appear. Hence, the following regular expression is used to extract C-C successions (marked in bold): $[\#|V]\mathbf{C}V\mathbf{C}[\#|V]$. Next, each consonant is assigned to one of the three major articulation place categories *labial*, *coronal* and *dorsal*. The succession counts are summarized in a quadratic matrix where the rows represent the preceding place of articulation and the columns the following place of articulation. Each matrix cell contains the number of times the respective place of articulation succession could be observed in the corpus. Subsequently, for each of the 9 possible successions a contingency table was created (Table 3).

	P_2	$\neg P_2$
P_1	$A: n(P_1 \rightarrow P_2)$	B: $n(P_1 \rightarrow \neg P_2)$
$\neg P_1$	$C: n(\neg P_1 \rightarrow P_2)$	$D: n(\neg P_1 \rightarrow \neg P_2)$

Table 3: Contingency table for the articulation place (P) succession from P_1 to P_2.

The succession counts were used to calculate ϕ coefficients, where A, B, C and D correspond to the four cells in Table 3.

$$\phi = \sqrt{\frac{\chi^2}{(A+B+C+D)}} \qquad (1)$$

The ϕ coefficient is a measure for the degree of association between two variables which can be derived from the fourfold χ^2 statistical significance test (see Rummel, 1970:298f for details). Sample ϕ values for the place of articulation successions of Egyptian Arabic can be seen in Table 4. A visual representation of the same matrix is provided in Figure 2. Note the at-a-glance analysis made possible by Figure 2 vs. Table 4.

	labial	coronal	dorsal
labial	−0.360	+0.187	+0.183
coronal	+0.259	−0.243	−0.068
dorsal	−0.010	+0.097	−0.121

Table 4: Matrix of ϕ values for Egyptian Arabic.

Figure 2 shows an example in which all diagonal values (self-successions of places of articulation) have negative associations. This tendency

[5] Note that they also included cases where the first and second consonant are part of a consonant cluster, which we ignored for our calculations. Furthermore, those languages where the number of consonant successions in the data was 20 or below were not included in our visualizations, thereby reducing the number of languages from about 4,500 to 3,200.

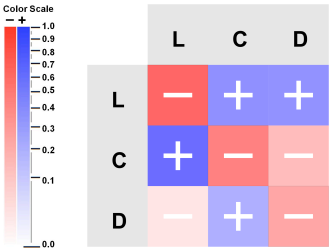

Figure 2: Visualization of the ϕ matrix from Table 4 (Egyptian Arabic), *L* stands for labial, *C* for coronal and *D* for dorsal. It can be seen that all diagonal values (successions of the same place of articulation) have negative associations (red color).

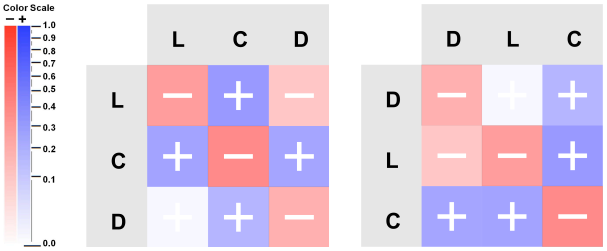

Figure 4: The ϕ matrix considering only the three main categories for all the data across languages. In the left figure, the categories are sorted according to their position in the oral cavity. In the right figure, the categories are sorted automatically, which shows that *D* and *L* are more similar to each other than *D* and *C*.

to alternate places of articulation can be observed in most languages and in the overall matrix visualizations including all data from all languages in the database (Figure 4).

4.2 General relations among places of articulation

As already mentioned, we tested whether it is useful to distinguish the two different subcategories dental (and alveolar) (*T*), and (alveo-)palatal (*C*). Figure 3 shows the resulting association values ϕ of place successions.

It can clearly be seen that *T* and *C* behave very similarly. A further interesting observation is that places of articulation tend to alternate (negative diagonal values for self-successions). As revealed in the succession graph of Figure 3, the places of articulation do not remain the same, but change to the closest alternative(s). In the case of *P* and *K* the closest distinct places of articulation (*T* and *C*) are preferred. In the case of *T* and *C*, however, this is somewhat different. Apparently, direct alternations between both are less probable. One plausible explanation could be that they are not distinct enough and thus either *K* or *P* are preferred as a following place of articulation, both having roughly the same distance. These observations led us to merge the places *T* and *C* in our further analyses and distinguish labial, coronal and dorsal consonants only, as in Figure 4.

Note that the cross pattern on the left in Figure 4, which now emerges very clearly, reinforces the hypothesis that the closest distinct place of articulation is preferred as successor.

4.3 Distribution across languages

Next, we examined the distribution of ϕ values for self-successions of places of articulation in about 3,200 languages. Self-successions correspond to the diagonal values of the ϕ matrices from the upper left to the lower right. As can be seen in the histogram in Figure 6, the peak of the distribution is clearly located in the area of negative association values. In the box-plots of Figure 5, which show the distributions for all three places of articulation separately, it is clearly visible that for each of the three places of articulation at least 75% of the languages included show negative associations. Furthermore, it can be seen that most outliers disappear when taking only the languages for which most data is available and thus statistics are more reliable. The same can be seen in the scatter plot in Figure 6, where the average ϕ value is always negative if the number of successions exceeds a certain threshold. For all three categories, the figures demonstrate that the same place of articulation is generally less frequently maintained than expected if there were no interdependencies between consonant co-occurrences.

5 Visualization of geo-spatial patterns

The most common approach to visually represent crosslinguistic information on areal (or genealogical) patterns is to put each language as a single pixel or a small icon to its location on a map. For instance, the WALS database (Haspelmath et al., 2005) includes 141 maps on diverse structural (phonological, grammatical, lexical) properties of languages. We transformed the results of our SPA statistics for each language in the ASJP database

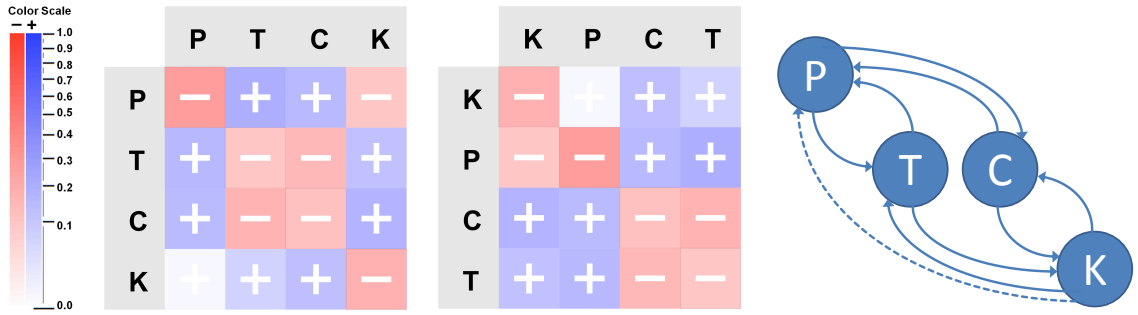

Figure 3: Successions of *P*, *T*, *C* and *K* in all languages. The "+" and "−" signs indicate the polarity of a succession (going from row to column category). The color saturation of the background indicates the strength of association. In the left figure, places of articulation are sorted according to their position in the oral cavity, in the middle figure an automatic similarity sorting of matrix rows and columns was applied. The right part of the figure shows an alternative view only on those successions that have a positive association.

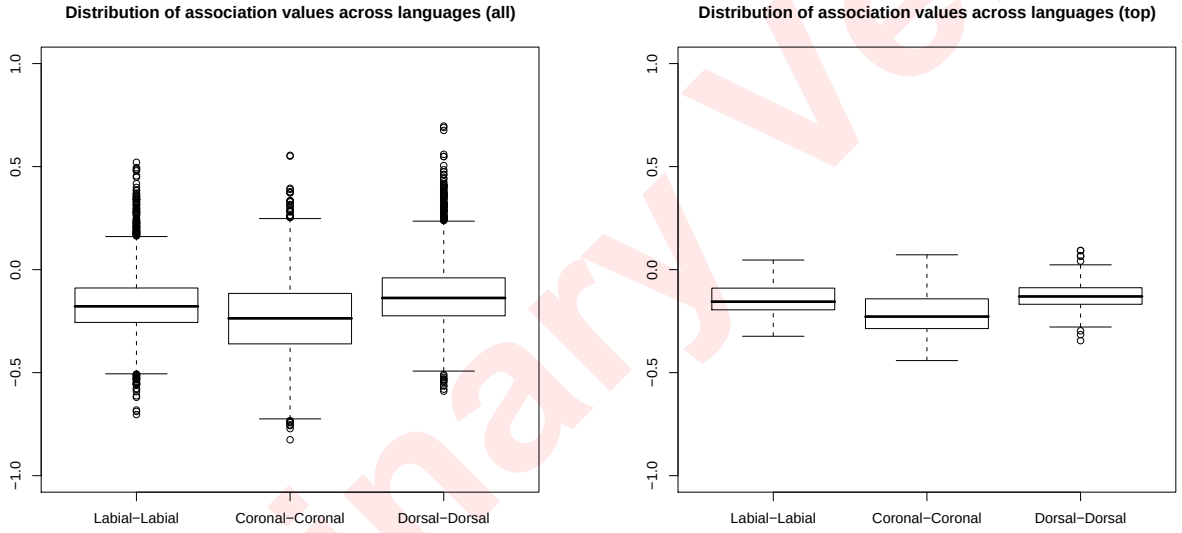

Figure 5: Boxplots showing the distribution of association strength values (ϕ) for self-successions of places of articulation. For the left boxplots about 3,200 languages were considered for which the Swadesh lists contained more than 20 successions. For the right boxplots only the top 99 languages were considered for which the Swadesh lists contained at least 100 successions, thereby removing most outliers and reducing the variance.

that is also included in the WALS database into a WALS map (Figure 7). The matrix visualization has been simplified in that the color of the icon represents the number of cells in the diagonal of the matrix whose value was below zero, i.e., the higher the number (0-3) the better the language conforms to SPA.

Some of the drawbacks of these maps include a high degree of overlap of data points in densely populated areas and the lack of correlation between information content and area size. In Figure 7, the fact that those languages with fewer negative diagonal cells are plotted on top of those with a higher number slightly distorts the overall picture that most languages adhere to the principle.[6] Besides that, the overall pattern in the densely populated areas is hardly visible, while sparsely populated areas waste space and hide the informational

[6]Likewise, the visualization would suggest to much adherence to the principle if those languages with more negative diagonal cells were plotted on top of those with fewer negative cells.

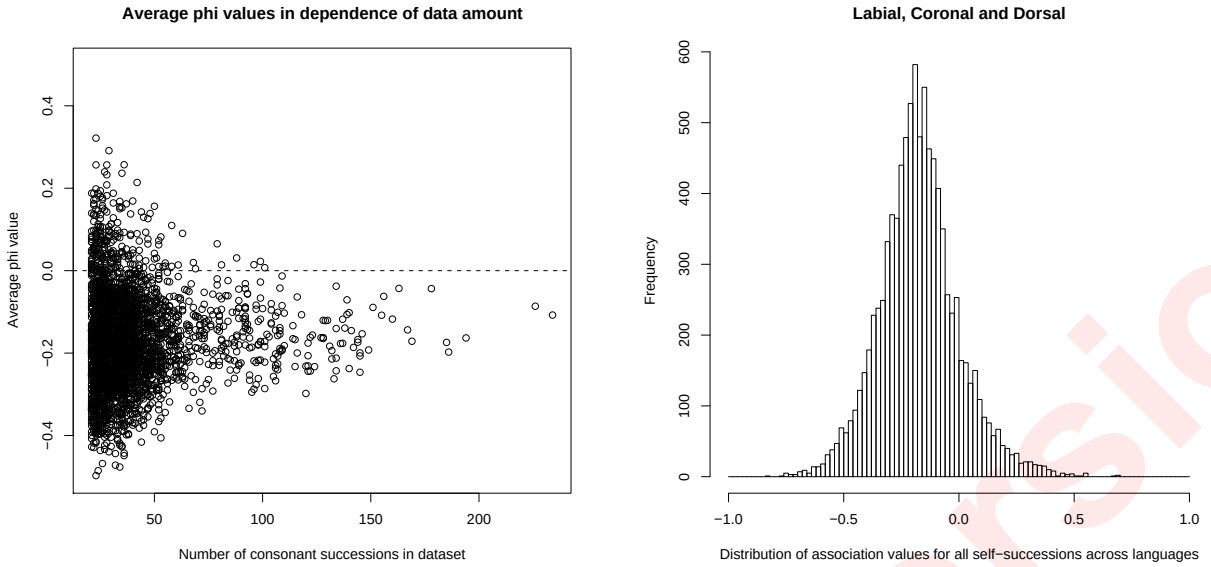

Figure 6: The scatter plot on the left displays the average ϕ values for self-successions of all places of articulation depending on the number of consonant successions (CVC) for each language in the sample. The histogram on the right shows the distribution of association strength values (ϕ) for self-successions of places of articulation in more than 3200 languages.

details. Finally, small clusters are difficult to find — they are not noticeable, and are sometimes even obscured by large clusters.

In order to avoid overlapping pixels we used a circular arrangement around the original location in the current analysis, taking the given ordering of elements into account (Bak et al., 2009a). The ordering usually corresponds to the coloring attribute starting with colors that occur least frequently. With this arrangement a natural looking visualization without artifacts is generated.

A way to obtain more space for regions with a high point density are Cartograms, which distort regions such that their size corresponds to a statistical attribute (Bak et al., 2009b; Tobler, 2004), in this case the number of languages in the database. The advantage is that more space is reserved to plot all important information on the map. In Figure 8, we show the density equalized distortion by cartograms and the overlap-free representation of the data points using pixel placement. Neighborhood relations and region shapes are at the same time maintained as accurately as possible in order to guarantee recognizability despite of distortion. The visualization reveals several clusters of non-conforming languages (marked with boxes). It remains for future work to investigate whether these clusters are an artifact of the database that we used or if they manifest an areal feature. Figure 8, in contrast to Figure 7, shows the 3,200 languages we investigated more closely and not just the ones included in WALS.

The representation thereby enables investigating spatial patterns free of hidden data and distributional biases.

6 Conclusions and future work

Our crosslinguistic investigation of SPA has confirmed the hypothesis that the phenomenon of Similar Place Avoidance is not a particular trait of Semitic languages, for which it was previously described, but is a linguistic universal tendency which can be observed in languages which are both genealogically and geographically unrelated. This can clearly be seen in the visualizations that display the conformity of each language in the database with respect to SPA. The overall picture for all languages not only shows that successive consonants with the same place of articulation tend to be avoided, but also that there is a tendency to avoid places of articulation that are too far away from the preceding place (cf. Figures 3 and 4).

We combine methods from statistics, NLP and Visual Analytics to provide a novel way of automatically assessing and visualizing linguistic features across a wide range of languages, thus al-

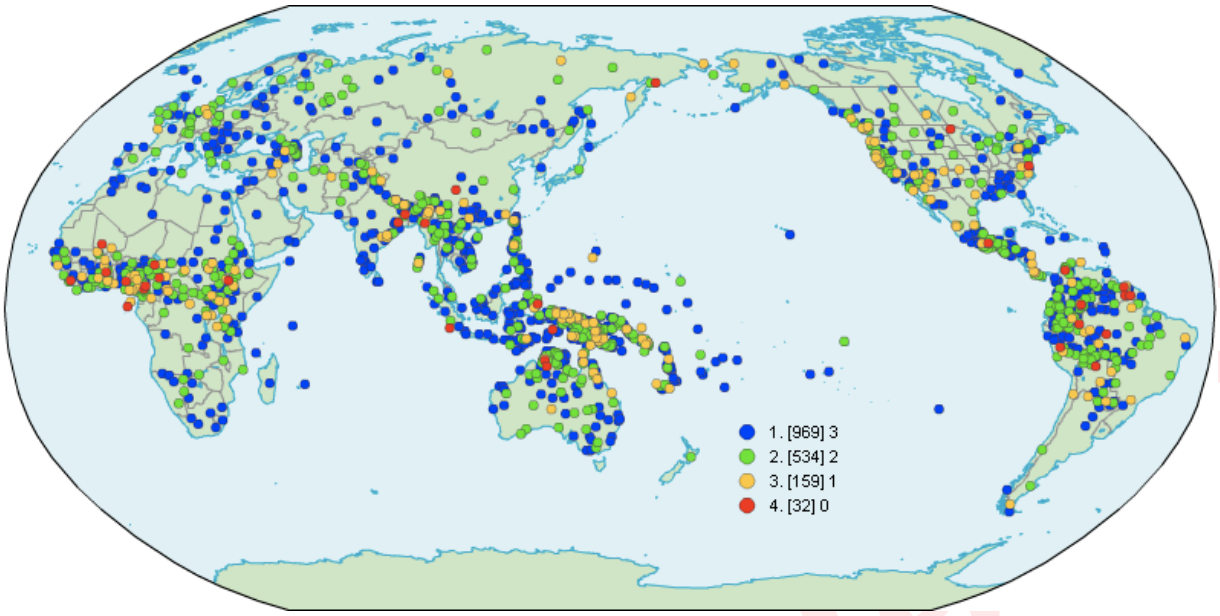

Figure 7: WALS map of the languages and their behavior with respect to SPA. The color indicates the number of self-succession ϕ values which are negative, i.e., which adhere to the SPA principle. Color mapping is from blue (conforming to SPA) to red. The numbers in square brackets indicate the number of languages in this group.

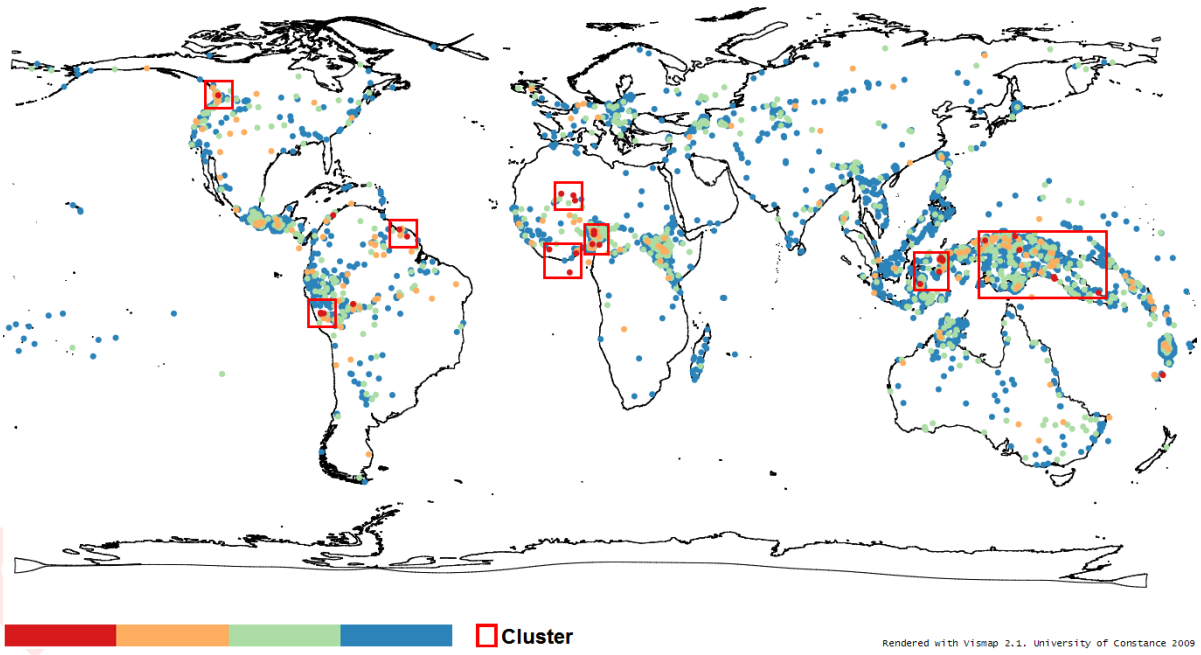

Figure 8: Density equalized distribution of the languages with respect to SPA. The area of the geographic regions corresponds to the number of languages in that location – represented by dots. Overlap is avoided using pixel-placement. The color mapping corresponds to the one used in the WALS map (Figure 7). Locations of nonconforming languages are highlighted with red boxes. Note that the number of languages in this map is about twice the number in the WALS map (7).

lowing for a gain of new insights and raising further interesting research questions that otherwise might easily go unrecognized.

With respect to SPA a more detailed exploration of the intricacies of phonological interdependencies is needed as part of our more widespread study of visually representing sound patterns in languages. As already hinted at in Pozdniakov and Segerer (2007), there are various other fascinating phenomena that are worth looking at, especially in regard to the interaction of vowels and consonants or vowel dependencies (such as vowel harmony) and consonant dependencies (such as SPA or consonant harmony). In particular, one could investigate why some languages apparently do not conform to SPA and if there is any co-variation to be uncovered between the adherence to the principle and other factors that might be interesting to explore and possibly reveal new insights into the structure of languages.

Acknowledgments

This work has been funded by the research initiative "Computational Analysis of Linguistic Development" at the University of Konstanz. The authors would like to thank Aditi Lahiri and two anonymous reviewers for valuable comments and suggestions.

References

Peter Bak, Florian Mansmann, Halldor Janetzko, and Daniel Keim. 2009a. Spatiotemporal analysis of sensor logs using growth ring maps. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):913–920.

Peter Bak, Matthias Schaefer, Andreas Stoffel, Daniel Keim, and Itzhak Omer. 2009b. Density equalizing distortion of large geographic point sets. *Journal of Cartographic and Geographic Information Science (CaGIS)*, 36(3):237–250.

Balthasar Bickel. in press. Absolute and statistical universals. In Patrick C. Hogan, editor, *The Cambridge Encyclopedia of the Language Sciences*. Cambridge: Cambridge University Press.

Paul Boersma and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32:45–86.

Paula Fikkert and Clara C. Levelt. 2010. How does place fall into place? The lexicon and emergent constraints in the developing phonological grammar. In Peter Avery, B. Elan Dresher, and Keren Rice, editors, *Contrast in Phonology: Perception and Acquisition*. Berlin: Mouton de Gruyter.

John Goldsmith. 1976. *Autosegmental phonology*. Ph.D. thesis, Massachusetts Institute of Technology.

Joseph H. Greenberg. 1950. The patterning of root morphemes in Semitic. *Word*, 6:161–182.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. 2005. The World Atlas of Language Structures Online. URL: `http://wals.info/`.

Gregory K. Iverson and Joseph C. Salmonts. 1992. The phonology of the Proto-Indo-European root structure constraint. *Lingua*, 87:293–320.

Daniel A. Keim, Florian Mansmann, Joern Schneidewind, Jim Thomas, and Hartmut Ziegler. 2008. Visual analytics: Scope and challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, Lecture Notes in Computer Science, pages 76–91. Springer.

Wiliam R. Leben. 1973. *Suprasegmental phonology*. Ph.D. thesis, Massachusetts Institute of Technology.

John J. McCarthy. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17:207–263.

Frans Plank. 1981. *Morphologische (Ir-)Regularitäten: Aspekte der Wortstrukturtheorie*. Tübingen: Gunter Narr Verlag.

Konstantin Pozdniakov and Guillaume Segerer. 2007. Similar Place Avoidance: A statistical universal. *Linguistic Typology*, 11(2):307–348.

Christian Rohrdantz, Thomas Mayer, Miriam Butt, Frans Plank, and Daniel A. Keim. 2010. Comparative visual analysis of cross-linguistic features. In *Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010)*, pages 27–32.

Rudolph J. Rummel. 1970. *Applied Factor Analysis*. Evanston, IL: Nortwestern University Press.

James J. Thomas and Kristin A. Cook. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr.

Waldo Tobler. 2004. Thirty five years of computer cartograms. *Association of American Geographer*, 94(1):58–73.

Søren Wichmann, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Matthias Urban, Sebastian Sauppe, Oleg Belyaev, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, and Helen Geyer. 2010. The ASJP database (version 12). URL: `http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm`.