Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
December 2010

Visualizing Vowel Harmony

Thomas Mayer, Christian Rohrdantz, Miriam Butt, Frans Plank, Daniel A. Keim

Submitted, February 2010
Revised, December 2010
Published by CSLI Publications

LiLT volume 4, issue 2 December 2010

Visualizing Vowel Harmony

THOMAS MAYER, CHRISTIAN ROHRDANTZ, MIRIAM BUTT, FRANS PLANK, DANIEL A. KEIM

Abstract

This paper deals with vowel harmony from a cross-linguistic perspective, with the aim of visualizing the distribution of vowels in corpora so that languages with vowel harmony can be distinguished from those lacking it. For this purpose vowel successions within words are statistically analyzed and visualized in a quadratic matrix whose rows and columns are ordered according to their distribution in the text, with more similar vowels occurring closer together. The method has been tested on the basis of Bible corpora in a variety of languages including well-known harmonic languages such as Turkish, Hungarian and Finnish as well as non-harmonic languages.

1 Introduction

This paper deals with vowel harmony (VH) from a cross-linguistic perspective, with the aim of investigating whether methods from Visual Analytics (Thomas and Cook 2005; Keim et al. 2008) can be used to aid linguistic analysis. Using available corpora in the form of Bible texts for 42 languages, we visualize the distribution of vowels in corpora so that languages with VH can be distinguished from those not containing it. We show that a visualization in terms of a quadratic matrix whose rows and columns are ordered according to a similarity measurement indeed allows for an at-a-glance visual analysis of VH patterns across languages. We also show that a data set of about 500

LiLT Volume 4, Issue 2, December 2010.
Visualizing Vowel Harmony.

types is sufficient for our analysis, thus providing a potentially useful tool for field linguists (Section 5).

With several case studies, we further demonstrate that our analysis is detailed enough to allow for an automatic visual identification of the precise types of harmony involved (see Section 6.3, case study on Udihe) and the identifcation of other types of phonological phenomena such as reduplication (see Section 6.2 on Warlpiri vs. Maori) and German umlauting (Section 6.4).

The work presented here is part of a wider ranging effort to visually represent cross-linguistic patterns (e.g., Mayer et al. 2010) with the ultimate goal of automatically extracting a phonological fingerprint for languages on the basis of corpora. The aim of visualizing patterns is to provide for a first, at-a-glance, mode of analysis that can function as a hypothesis generating device which supports linguistic analysis. The phenomenon of VH is a test case to see whether such visualizations can indeed live up to our expectations. Hence the goal of our visualizations is not to account for all intricacies of the phenomenon at hand, but to reveal general patterns that can be subjected to further detailed linguistic analysis.

A language is considered to have VH when the co-occurrence of vowels within a certain domain (usually the phonological word) is constrained with respect to certain language-specific features which each vowel within the domain has to share (see van der Hulst and van de Weijer 1995 for an overview). In most cases VH can be conceived of as spreading from left to right, with each non-initial vowel taking over the harmonic feature of the previous vowel. This results in co-occurrence restrictions on the vowels within stems (stem-internal harmony) and beyond stems, with stem and affix vowels limited to members of the relevant harmony classes. For present purposes stem-internal and stem-external harmony are not distinguished. Furthermore, no distinction is made in our calculations between left-to-right and right-to-left harmony spreading. Thus, umlaut-type processes as in Germanic, where certain vowels in suffixes alter the vowel of stems (typically only one vowel, in contradistinction to typical VH, affecting all stem vowels), are subsumed under the notion of harmony. For the purpose of this investigation both are treated the same in that only whole word forms are taken into consideration, independent of whether they are morphologically complex or not.

The present study is corpus-based, and this necessitates a limitation of the domain to what is marked off through spaces in the orthography of the language. Therefore languages where VH is active across word boundaries (as marked in the orthography) cannot be detected with our method. In Chamorro, for instance, a type of VH called vowel fronting in the first vowel of the stem is triggered by particles that are mostly written as separate words (see Topping 1980) and thus would not contribute to the relevant vowel successions in our calculations. For the same reason only languages

whose orthography is close to the actual pronunciation of the words could be included in the sample. One-to-many or many-to-one relations from orthography to phonology (with regard to vowels in particular) add too much noise to the calculations and lead to useless results.[1] The same holds for languages with spelling systems where vowel and consonant pairs are represented by one symbol, such as in the Russian alphabet where Я/я, for instance, denotes the sequence of a glide /j/ plus the vowel /a/. Nevertheless, preliminary explorations showed that interesting findings could be obtained from corpora of such languages (see below the case of German umlaut) so that we decided to include some of them in the overview of our results.[2]

Languages differ with respect to how many and which features are involved in harmony. As will be seen below, Turkish has both palatal (front/back) and a restricted type of labial (rounded/unrounded) harmony, whereas Finnish only has palatal harmony. Other harmonizing features that can be found in the languages of the world include [ATR], [RTR], [LOW], [HIGH]. We deal equally with all such features, rather than distinguishing harmony systems in terms of the features involved.

We experimented with a statistical analysis of vowel successions in word forms of the languages under investigation. For this purpose, vowels and consonants are separated from each other and are represented on different levels (or tiers), as in the framework of autosegmental phonology (Goldsmith 1976). That is, a non-local phenomenon such as VH, where a harmony feature can spread across a number of consonants, can be treated as a local assimilation process on the vowel tier. In this study, VH is considered to be local in this sense, with features spreading from one vowel to the other.

There are several reasons why an analysis of vowel successions for VH has to be robust with regard to non-harmonic vowel transitions. First, most VH systems allow occasional disharmony in their stems and affixes. Disharmonic stems are mostly high-frequent lexemes in the language or borrowings from other (non-harmonic) languages that failed to adapt to the harmony system of the recipient language. Disharmonic affixes usually reflect the transition from clitic to affix status in that they do not undergo harmony or simply constitute exceptions to the general harmony rules. Second, some languages have vow-

[1] In English, for example, the letter <e> reflects several phonemes in words like *red, precede, belong* whereas the phoneme /i/ is written with different letters in the words *feel, be, clear*. Even worse are cases where one letter stands for both vowels and consonants depending on the context, e.g., the letter <y> in English can denote the vowel /i/ as in *lady* or the consonant /j/ as in *yoghurt*.

[2] Although we already achieve good results with regard to VH despite the limitations detailed above, it would of course be more desirable to perform the analysis on a phonological transcription, in which case a cross-linguistic comparison of vowels would also be possible. At present we are unable to conduct such an experiment due to the lack of a sufficient amount of cross-linguistic data.

els that do not conform to the harmony in some way or another because they are either not specified for the harmony feature or lack the relevant counterpart for the respective feature value. Such vowels are usually called *neutral* vowels. Neutral vowels are distinguished in terms of whether they do not participate in the harmony at all (in which case they are called *transparent* vowels because they act as if they were invisible to the following vowel) or whether they block the preceding harmony class and impose their own value to the following vowels. The latter type of neutral vowels is called *opaque*.

The following section introduces two well-known languages, Turkish and Finnish, to illustrate the basics of VH in more detail and to justify the approach of vowel successions as the basis for detecting VH patterns. In Section 3 a brief overview of related work is given. Section 4 describes the technique for analyzing vowel successions and visualizing their results. Questions with regard to how much data is needed to produce stable results are discussed in Section 5. Section 6 discusses a few case studies of VH languages, to highlight strengths and difficulties of the present approach. In particular, we look at detecting VH in "new" languages as well as visually differentiating VH patterns from other phonological phenomena such as reduplication or umlauting. Section 7 wraps up with a few general conclusions.

2 Vowel harmony systems and relevant parameters for cross-linguistic variation

In this section two harmony systems from Turkish and Finnish are described in more detail in order to exemplify the basic concepts of VH and the parameters on which languages can differ. Turkish has two harmony processes that are active at the same time, whereas Finnish VH exemplifies the concept of neutral vowels.

The following outline of Turkish and Finnish simplifies these VH systems and only covers the more straightforward regularities. Disharmonies and neutral vowels are largely ignored here; for further detail the interested reader is referred to specialist literature such as Clements and Sezer (1982) for Turkish and Ringen and Heinämäki (1999) for Finnish.

2.1 Turkish

Turkish vowels can be classified according to the features [±HIGH], [±FRONT], [±ROUNDED], as shown in Table 1.

With regard to their behavior in VH processes Turkish vowels can be divided into three clusters. The first cluster contains only [−HIGH,−ROUNDED] vowels /e a/ whereas the second cluster contains [+HIGH] vowels /i ü ı u/. These vowels are the targets for the harmony processes, while the remaining non-high rounded vowels /ö o/ do not participate as targets at all, which means that they either only occur in disharmonic words or in initial position.

	Front		**Back**	
	Unrounded	**Rounded**	**Unrounded**	**Rounded**
High	i	ü	ı	u
Low	e	ö	a	o

TABLE 1 Turkish vowels

The target clusters in turn can be related to the vowels that trigger the respective harmony pattern.

Turkish suffixes show alternations with respect to two harmony processes called palatal and labial harmony. For palatal harmony each vowel is characterized by its frontness or backness [±FRONT] and triggers the vowel /e/ for [+FRONT] and /a/ for [+BACK]. For instance, the noun *adam* 'man' shows the plural form *adam-lar* with the back vowel, whereas the plural of *ev* 'house' is *ev-ler*. The same holds for all other vowels that are not identical to the suffix vowel, e.g., *göz* 'eye' takes the plural form *göz-ler* because the stem vowel is specified for [+FRONT], whereas *top* 'ball' gets the plural form *top-lar*. Their [FRONT, ROUND] values are filled in as determined by the last vowel of the stem. Similarly, vowels of suffixes showing labial harmony are only specified for the feature [+HIGH]. Their [±FRONT, ±ROUND] values are specified by the last vowel of the stem. Depending on the type of harmony involved the following transitions from one vowel to another are allowed (see Table 2). Note that the vowels /o,ö/ do not occur on the right side of the table.

		palatal harmony	labial harmony
a	→	a	ı
ı	→	a	ı
u	→	a	u
o	→	a	u
ö	→	e	ü
ü	→	e	ü
i	→	e	i
e	→	e	i

TABLE 2 Turkish harmony classes

Since the harmony of subsequent suffixes is usually dependent on the values of the preceding suffix, harmony of either type spreads from left to right from one vowel to the other. This is seen with a case inflected noun in the plural. In Turkish nouns the plural marker *-lAr*[3] is followed by the case marker (which is zero in the nominative as in the above examples where only *-lAr* oc-

[3]Unspecified vowels are usually transcribed with capital letters.

curs). The plural shows palatal harmony (i.e., a two-way contrast between *-ler* and *-lar*) whereas some case suffixes show labial harmony (i.e., a four-way contrast as with the genitive *-In*: *-in*, *-ın*, *-un*, *-ün*). The genitive suffix vowel in the singular form *top-un* where it is triggered by the preceding stem vowel changes to /ı/ in the plural form *top-lar-ın* when the plural marker stands in between the stem and the genitive suffix (see Table 3).

Feature specification	Genitive suffix
[+FRONT,−ROUND]:	deniz-in, ev-in
[+FRONT,+ROUND]:	tütün-**ün**, çöl-**ün**
[−FRONT,−ROUND]:	kadın-ın, adam-ın
[−FRONT,+ROUND]:	sabun-**un**, top-**un**

Feature specification	Genitive suffix with plural suffix
[+FRONT,−ROUND]:	deniz-ler-in, ev-ler-in
[+FRONT,+ROUND]:	tütün-ler-**in**, çöl-ler-**in**
[−FRONT,−ROUND]:	kadın-lar-ın, adam-lar-ın
[−FRONT,+ROUND]:	sabun-lar-**ın**, top-lar-**ın**

TABLE 3 Turkish harmony spreading from left to right. The intervening plural suffix *-lAr* imposes the specification of the genitive suffix vowel as to its rounding feature. Notice the different realization of the genitive suffix *-In* with and without the plural suffix.

VH as in Turkish should therefore be detectable when looking at local dependencies between adjacent vowels within a word. If we only look at vowel transitions, the following vowel-to-vowel transition matrix (Table 4) emerges, where a plus marks a permissible transition and a minus marks a non-permissible transition in terms of either palatal or labial harmony.

	a	ı	u	o	ö	ü	i	e
a	+	+	−	−	−	−	−	−
ı	+	+	−	−	−	−	−	−
u	+	−	+	−	−	−	−	−
o	+	−	+	−	−	−	−	−
ö	−	−	−	−	−	+	−	+
ü	−	−	−	−	−	+	−	+
i	−	−	−	−	−	−	+	+
e	−	−	−	−	−	−	+	+

TABLE 4 Permissible vowel successions in Turkish harmonic words. The rows mark the preceding vowels whereas the columns mark their successors.

Our aim is to arrive at a visualization, also in matrix form, of the vowel succession patterns. This matrix should be generated automatically from corpus data without specialized knowledge about the language under consideration. The harmony classes will be detectable in the saturation and coloring of the cells as well as the ordering of the vowels that make up the visualization.

2.2 Finnish

Finnish vowels can be characterized by the features in Table 5. In Finnish VH the initial syllable of each simple (non-compound) word controls the frontness or backness of the suffix vowels. If they occur non-initially, the neutral vowels are transparent and therefore do not affect the VH. If the stem to which the suffixes are attached only contains neutral vowels, these (usually) behave as front vowels with respect to the following suffix vowels (Ringen and Heinämäki 1999, p. 306), e.g., *vélje-llA* → *véljellä* (brother-ADESSIVE).

	Front	**Neutral**	**Back**
Open	ä		a
Mid	ö	e	o
Closed	y	i	u

TABLE 5 Finnish vowels (note that there is no neutral open vowel)

trigger		permissible following vowels
ä	→	ä, ö, y, e, i
ö	→	ä, ö, y, e, i
y	→	ä, ö, y, e, i
e	→	ä, ö, y, e, i, a, o, u
i	→	ä, ö, y, e, i, a, o, u
a	→	e, i, a, o, u
o	→	e, i, a, o, u
u	→	e, i, a, o, u

TABLE 6 Finnish harmony

Unlike Turkish, there is only one harmony system (palatal harmony) active in the language. Therefore, suffix vowels do not change their quality through intervening suffixes that are subject to a different harmony pattern as with Turkish case suffixes with an intervening plural (as in Table 3), which can change the vowel of the genitive marker.

However, palatal harmony in Finnish can be interrupted by the neutral vowels /e, i/, in which case successive vowels do not share the [±FRONT] value. As can be seen in Table 6, /e, i/ are permissible after any other vowel

because they are invisible to VH in non-initial position. For this reason the rows and columns for /e, i/ in the matrix in Table 7 are all filled with pluses.

	a	ä	e	i	o	ö	u	y
a	+	−	+	+	+	−	+	−
ä	−	+	+	+	−	+	−	+
e	+	+	+	+	+	+	+	+
i	+	+	+	+	+	+	+	+
o	+	−	+	+	+	−	+	−
ö	−	+	+	+	−	+	−	+
u	+	−	+	+	+	−	+	−
y	−	+	+	+	−	+	−	+

TABLE 7 Permissible vowel transitions in Finnish harmonic words

2.3 Non-harmonic sequences

Looking only at local transitions within words one encounters several difficulties. Neutral vowels as found in Finnish confuse the harmony calculations because they do not conform to the restriction that only certain vowels can be found next to each other. Apart from this, disharmonic stems and suffixes also contribute to a less clear-cut distinction between harmony classes. In the Turkish progressive suffix *-Iyor*, for instance, only the first vowel harmonizes, whereas the second vowel is invariant.[4] In both Turkish and Finnish, compound words are considered as two distinct words so that suffixes adopt the harmony values of the last member of the compound. Since compounds are written without internal spaces this is another cause for the fuzzy boundaries in the visualization matrices as described in the next sections. However, over a considerable amount of data (see Section 5) languages with vowel harmonies should have a skewed distribution as to the vowel sequences, whereas non-harmonic languages should have a more or less balanced distribution of their vowels.[5]

3 Related work

The purpose of visualizing information is to represent abstract data so that interesting patterns and relations become visible that otherwise might easily go unrecognized in the sea of detail. Such visualizations play a crucial role

[4] With respect to following vowels the progressive suffix vowel /o/ acts like an opaque vowel.

[5] This does not hold for positional distribution, especially with respect to word-final vowel occurrences. Note that vowel successions without an intervening consonant are not taken into consideration because it is hard to tell in each case whether they constitute a diphthong or a case of hiatus.

in the discipline of Visual Analytics (Thomas and Cook 2005; Keim et al. 2008) that integrates both automatic and interactive visual data analyses. It offers methods and techniques to conduct a visual data exploration in order to discover interesting patterns and relations, generate hypotheses and confirm or reject previous assumptions about the data at hand.

Early approaches to the visualization of text data have mainly focused on providing topical overviews of document collections (Honkela et al. 1995; Wise et al. 1999; Wise 1999) as well as detailed topical insight into document collections (Hearst 1995; Fekete and Dufournaud 2000; Havre et al. 2002; Don et al. 2007; Collins et al. 2009; Strobelt et al. 2009), and are mostly related to the field of Information Retrieval. Since then, many related visualizations have evolved. Some of the recent approaches are based on much more sophisticated text processing methods which also involve some linguistic knowledge. Approaches that are concerned with the visual analysis of affective content in large document collections (e.g., news articles, weblog entries or customer reviews) have appeared recently in the field of opinion and sentiment analysis (Gregory et al. 2007; Gamon et al. 2008; Wanner et al. 2009; Oelke et al. 2009).

While all of the work mentioned primarily deals with extracting and visualizing the topical content of text data, the visual exploration of other text features and natural language phenomena has received little attention. Honkela et al. (1997) have obtained visual syntactic category clusters by generating self-organizing maps based on word context vectors. Keim and Oelke (2007) have extracted and visualized diverse statistical text properties on different hierarchy levels for literature analysis and authorship attribution, and Abbasi and Chen (2007) have extracted and visualized detailed text features to enable a visual classification of documents that is not only based on topic content but also on style and sentiment. Wattenberg and Viégas (2008) created the *Word Tree* visualization that was primarily aimed at visualizing the content structure of texts, but which can also be used to visualize language features as shown by the example of a tree containing Greek nominal suffixes. Recently, Albrecht et al. (2009) introduced an interactive tool for the correction of erroneous machine translation output with visual components.

As to the identification of VH, there have been several earlier approaches that also try to capture harmonic patterns from raw texts (Hare 1990 on Hungarian; Ellison 1994 for Turkish and Yoruba). Apart from that Altmann (1986) derives a statistical method to calculate the tendency for languages to have identical vowels occurring next to each other (what he calls *Tendenzielle Vokalharmonie*).

More recently, Goldsmith and Riggle (2007) have investigated VH in Finnish by taking advantage of information theoretic concepts in order to better understand the phonological structure of the language. Their work

partly differs from our approach in that they concentrate on developing a device that is able to quantitatively defend a VH analysis for Finnish on the basis of a corpus. In Goldsmith and Xanthos (2009), a spectral approach for the identification of VH classes in Finnish is described. Their method results in a classification where front and neutral vowels are in a single cluster whereas back vowels form a group of their own. Both these approaches have been applied to a single language (Finnish), but could in principle be used for other languages as well. Baker (2009) is an extension of Goldsmith and Riggle (2007), which attempts to detect VH in four different languages (Turkish, Finnish, English and Italian) by examining two methods for learning and modeling VH on the basis of text corpora (expectation maximization with Hidden Markov Models and pointwise mutual information in a Boltzmann distribution). These methods correctly predict that Turkish and Finnish have VH and at the same time do not find harmony in English or Italian.

Another work which particularly aims at comparing different languages with respect to their VH tendency is Harrison et al. (2004). Their online harmony calculator is a tool for the quantification of vowel co-occurrences in a corpus. It determines the percentage of harmonic words in the corpus and the harmony index, i.e., the extent to which the percentage exceeds random chance. In order to achieve this, a unique harmony threshold for each corpus is calculated, which represents the percentage of words that are expected to be harmonic purely by chance.

However, to the best of our knowledge there are no published approaches for the visualization and visual analysis of linguistic phenomena like VH and no work trying to visually compare such linguistic properties across a larger set of languages.

4 Technique

4.1 Data gathering and processing

As already mentioned, the goal of the present investigation is to detect VH (or similar phenomena) by statistically analyzing vowel successions. For this purpose a reliable large-scale language resource in digitalized form is required for a variety of languages. Therefore, it is not appropriate to just take a list of citation forms of words as listed in dictionaries. These in general do not contain inflected word forms where VH will crucially be seen in action. A corpus of running text is required for this purpose. The Bible is a suitable text because it is available in many languages, and matters of content are irrelevant, with our interest focused exclusively on form rather than on meaning.

We have therefore collected Bible texts (whole New Testament or only Bible portions) for different languages. From each Bible text we used a list of all types (i.e., all word forms occurring in the corpus). Basing the exploration

of VH on the list of types (rather than tokens) guarantees that each word is considered only once in the applied statistics. This has the desired effect that highly frequent words do not dominate the result. Especially the very frequently repeated proper names could otherwise have strongly biased the outcome, as in many Bible translations they were not adapted to the phonology of the recipient language or at least not according to its common vowel patterns.

The digitalized Bible texts not only serve as a source for the word forms of a language but also for its vowels. A list of vowels for each language has been automatically extracted with the help of Sukhotin's algorithm for vowel/consonant discrimination (Sukhotin 1962), whose results have been manually revised in order to avoid errors.

4.2 Counting vowel successions

In a first step, an elementary statistical processing is performed for each language, simply summing up the vowel successions occurring in all the types. For this purpose, we define a vowel succession as an ordered pair of vowels within a word. Sequences of two vowels without intervening consonants have been ignored: in the vast majority of instances these will be diphthongs; less desirably, some heterosyllabic vowels — which could potentially harmonize — will also have been eliminated as a result of this analytic decision. To give an example, the word *harmonic* would contribute to the count of the vowel succession "o follows a" which we will refer to as (a-▷o) and to the count of the vowel succession (o-▷i).

As a result of summing up the vowel successions of each type within the Bible, a matrix of succession counts is obtained. An example for such a succession matrix is given in Table 8.

	a	ä	e	i	o	ö	u	y
a	3548	20	1940	1893	831	0	944	24
ä	35	944	806	820	10	138	33	266
e	1623	1144	1495	1608	419	56	497	187
i	1580	854	1514	1044	376	46	355	135
o	1384	7	1032	902	284	0	294	8
ö	7	125	54	39	0	3	1	18
u	1464	6	1085	850	315	1	547	8
y	39	656	368	368	35	75	4	251

TABLE 8 Example of a matrix with succession counts for the Finnish Bible. The successions go from the row letter to the column letter. The succession (a-▷e), for instance, occurred 1940 times.

4.3 Statistics

The simple matrix with the counts of vowel successions provides a rather general overview. Some high or low values are salient and usually it can be seen that some vowels appear with a much higher overall frequency than others. For most languages the strong variance between the overall frequencies of distinct vowels is the dominating effect visible in the matrix.

To add detail, the succession probabilities are calculated. That means that for each vowel it is calculated with which probability certain other vowels are expected to be observed next. Of course, highly-frequent vowels in most cases still have a higher probability of succeeding any other vowel than low-frequent vowels. This is easily seen in the example for Finnish (see Table 9): After any vowel it is much more probable to observe an /e/ than to observe an /ö/, which can be deduced from the corresponding matrix columns. While these large-scale differences stand out, small-scale differences can be of even more interest. They possibly indicate phenomena which can barely be detected when skimming over a text. For instance, infrequent vowels generally have a low probability of succeeding other vowels and therefore rather small variations in their absolute probabilities may already be interesting in a statistical sense.

	a	ä	e	i	o	ö	u	y
a	0.386	0.002	0.211	0.206	0.090	0.000	0.103	0.003
ä	0.011	0.309	0.264	0.269	0.003	0.045	0.011	0.087
e	0.231	0.163	0.213	0.229	0.060	0.008	0.071	0.027
i	0.268	0.145	0.256	0.177	0.064	0.008	0.060	0.023
o	0.354	0.002	0.264	0.231	0.073	0.000	0.075	0.002
ö	0.028	0.506	0.219	0.158	0.000	0.012	0.004	0.073
u	0.342	0.001	0.254	0.199	0.074	0.000	0.128	0.002
y	0.022	0.365	0.205	0.205	0.019	0.042	0.002	0.140

TABLE 9 Example of a matrix with succession probabilities for the Finnish Bible. Again, the successions go from the row letter to the column letter. The vowel /a/ for instance was followed by the vowel /e/ in 21.1% of the cases. For display purposes all numbers have been rounded to three decimals. Without rounding the row sums are equal to 1.

This lead us to experiment with several tests for dependence (χ^2, t-test, likelihood-ratio, pointwise mutual information; see Manning and Schütze 1999) for vowels. Since our aim is to compare different languages on the basis of corpora of different sizes, the visualizations have to be based on normalized values in order to be able to compare the results visually. The only reasonable normalization method that we could find for the several tests for

dependence is the association strength value ϕ, which is a normalized form of χ^2 and will be explained in more detail further on.

	e	not(e)
a	A = 1940	B = 7260
not(a)	C = 6354	D = 19861

TABLE 10 Example of the fourfold (contingency) matrix for the succession (a-▷e) in Finnish. The expression "not(a)" stands for the set of all vowels except /a/ and the same with "not(e)". Note that the four cells of the matrix have names (A, B, C and D) that are important for the Formulas 1.1 and 1.2.

To get a value of how much a vowel succession deviates from its expected value the fourfold χ^2 formula (see Formula 1.1, Manning and Schütze 1999) is applied. The higher the values, the more significant in a statistical sense is the deviation of observed frequencies from expected frequencies. The test quantifies the influence of the independent variable (e.g., /a/ in Table 10) on the dependent variable (e.g., /e/ in Table 10).

$$\chi^2 = \frac{(A+B+C+D)\cdot(A\cdot D - C\cdot B)^2}{(A+C)\cdot(B+D)\cdot(A+B)\cdot(C+D)} \tag{1.1}$$

The χ^2 value depends on the sample size and therefore is not easily interpretable and comparable among sets of different size. To overcome this problem the correlation coefficient ϕ was applied (see Formula 1.2, Manning and Schütze 1999).

$$\phi = \sqrt{\frac{\chi^2}{(A+B+C+D)}} \tag{1.2}$$

The ϕ coefficient represents the association strength and, when derived from the χ^2 values, always lies in the interval between 0 and 1.[6] The matrix of ϕ values for Finnish can be seen in Table 11.

One piece of information that is not contained in the matrix of the ϕ values is if a vowel succession occurs more or less frequently than expected. In order to find this out, the probability value for every vowel succession is considered, e.g., for the Finnish matrix (Table 9) and the succession (a-▷e) the value is 21.1%. Then, for the succeeding vowel (here /e/) the overall succession probability is calculated, which is independent of the first vowel. No matter which vowel is observed first, in 23.4% of the cases /e/ is the next

[6] When calculated directly from the fourfold matrix the ϕ values lie between -1 and +1, where a negative sign indicates a negative association among the two binary variables.

	a	ä	e	i	o	ö	u	y
a	0.149	0.200	0.033	0.010	0.063	0.056	0.061	0.086
ä	0.180	0.203	0.022	0.042	0.076	0.118	0.075	0.121
e	0.047	0.092	0.025	0.020	0.009	0.005	0.009	0.004
i	0.006	0.056	0.023	0.039	0.001	0.006	0.026	0.007
o	0.064	0.119	0.025	0.016	0.012	0.034	0.000	0.052
ö	0.046	0.109	0.003	0.011	0.022	0.003	0.023	0.025
u	0.057	0.126	0.017	0.012	0.014	0.034	0.073	0.055
y	0.130	0.195	0.016	0.004	0.042	0.080	0.064	0.168

TABLE 11 The matrix with the ϕ effect strength values for Finnish. For display purposes all numbers have been rounded to three decimals.

	a	ä	e	i	o	ö	u	y
a	+	−	−	−	+	−	+	−
ä	−	+	+	+	−	+	−	+
e	−	+	−	+	+	−	+	−
i	−	+	+	−	+	−	−	−
o	+	−	−	−	+	−	+	−
ö	−	+	−	−	−	+	−	+
u	+	−	−	−	−	−	+	−
y	−	+	+	−	−	+	−	+

TABLE 12 The matrix shows whether a succession occurred more (+) or less (−) frequently than expected. Compare this with the expected result in Table 7.

vowel. So the observed frequency with 21.1% of the cases is lower than expected (23.4%). Table 12 shows a matrix where a "−" is displayed whenever a succession occurs less frequently than expected and a "+" whenever it can be observed with a higher frequency than expected.

Since our calculations are only a preliminary step for the visualizations, the χ^2 (or ϕ) values are not tested for their significance. Rather, they are used as the input for the visualizations where the user can decide for themselves whether a certain value for a vowel succession is considered to be significant (not in the statistical sense, but in the eye of the beholder) for a certain pattern to stand out. After all, the statistical significance thresholds are an arbitrary agreement on when something is considered to be significant. In addition, statistical significance heavily depends on the amount of data taken into consideration. Our tests have shown that the statistical significance values (likelihood ratio, t-test, χ^2) tend to increase with more data, making it more difficult to compare their values across languages of different sample sizes. The effect strength value ϕ, however, does not depend on the amount of data as long

as the given sample is representative (see Section 5). The task of the visual component is to devise the visualization in a way that the decision whether a given pattern is salient can be made by the user.

4.4 Visualization & visual analysis

Reading and comparing numerical matrix entries is a slow and exhausting task for analysts and the discovery of previously unknown relations is hard to achieve. In contrast, representing abstract data as an information visualization by mapping numeric values to suitable pre-attentive visual variables can boost the analysis process. Analysts can detect interesting patterns without any cognitive overload. Useful interactions then allow a further exploration that is more focused and provides details on demand. The discipline of Visual Analytics offers methods and techniques to conduct a visual data exploration in order to generate hypotheses and confirm or reject previous assumptions about the data.

	a	ı	u	o	ö	ü	i	e
a	0.27	0.43	-0.14	-0.06	0.02	-0.13	-0.26	-0.28
ı	0.16	0.29	-0.11	0.08	-0.01	-0.08	-0.19	-0.19
u	0.13	-0.14	0.46	0.02	-0.00	-0.05	-0.14	-0.14
o	0.07	-0.11	0.43	-0.02	0.01	-0.05	-0.10	-0.11
ö	-0.11	-0.09	-0.05	-0.03	0.01	0.37	-0.09	0.16
ü	-0.12	-0.11	-0.06	0.01	-0.01	0.51	-0.11	0.13
i	-0.20	-0.22	-0.12	0.07	-0.00	-0.09	0.32	0.21
e	-0.26	-0.25	-0.13	-0.06	-0.01	-0.10	0.40	0.28

TABLE 13 The matrix of ϕ values for Turkish.

	a	i	o	e	u
a	-0.003	-0.075	0.094	-0.025	-0.018
i	-0.025	-0.004	0.064	-0.036	0.005
o	-0.028	-0.006	-0.075	0.098	0.026
e	-0,001	0.063	-0.073	0.016	0.021
u	0.077	0.038	-0.036	-0.057	-0.043

TABLE 14 The matrix of ϕ values for Spanish.

The matrices of ϕ values in Table 13 and Table 14 reveal the dependence of following vowels on preceding vowels and potentially enable linguists to determine whether a language has VH or other interesting patterns of vowel distribution. However, as the matrices for Turkish (contains VH) and Spanish

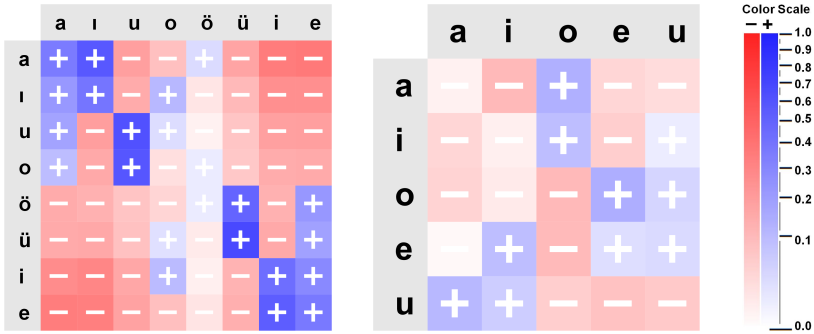

FIGURE 1 The visualized ϕ matrix for Turkish (left) and Spanish (right)

(contains no VH) also show, picking out patterns among a forest of numbers is not an easy task for most humans. In contrast, a corresponding visualization which allows an at-a-glance analysis is shown in Figure 1.

An intuitive way to visualize the matrices is to map the numerical values to different color shades, with color being a powerful and widely used visual variable (Wyszecki and Stiles 2000). Before visualization, however, a data transformation step is to be recommended: rows and columns of the matrices have to be sorted according to similarity in order to render a better visibility of patterns, as motivated in Bertin (1983) and Henry and Fekete (2006).

Matrix arrangement

To make relations between vowels of similar behavior visible, it is essential to sort the rows and columns of the matrices in a meaningful way. Only if a certain pixel coherence can be guaranteed will vowel succession patterns become evident. To enable a sorting of vowels, a first step is to calculate the numerical dissimilarity between vowels. To do so, for each vowel a feature vector is created that corresponds to the ϕ-values of its matrix row (with algebraic signs). Next, a distance function between feature vectors has to be defined that quantifies the dissimilarity of the ϕ-values of two vowels at a time. Different distance functions were tested and the one that yielded the best results can be found in Formula 1.3. If the entries contained by two feature vectors at the same index position have different algebraic signs, then a distance value of 1 is summed up. If both entries have the same sign, the square of their difference is summed up. As this difference necessarily is contained in the interval [0,1] the squaring decreases the summand. Thus, pairs of vectors containing different signs at the same index are considered dissimilar.

$$dist(x,z) = \sum^{n}{}_{i=1} d(x_i, z_i), \tag{1.3}$$

$$\text{where } d(x_i, z_i) = \begin{cases} 1 & \text{if } sign(x_i) \neq sign(z_i), \\ (x_i - z_i)^2 & \text{else.} \end{cases}$$

The distance measure in Formula 1.3 is then used in the sorting process. The first row in the matrix of any language is fixed as the row belonging to the vowel with the smallest Unicode value (usually the vowel [a]). A nearest neighbor sorting is done next: The most similar row (vector in high-dimensional feature space) to the /a/-vector is searched for and the corresponding vowel is placed in the second position. Next, the most similar vector to this second vowel is identified among the remaining ones. This procedure is iteratively repeated until there is no vowel left.

After sorting the rows, the vowel columns are sorted in exactly the same order. We also tried to sort columns and rows independently but came to the conclusion that this was not desirable as the diagonal of the matrix lost its general meaning (self-successions). Our tests showed that having the same row and column order is an important visual cue that helps in understanding the matrix and is more beneficial for the analysis process than an independent sorting of rows and columns.

Data mapping

For the visualization of the matrix containing absolute counts of vowel successions all values had to be normalized. This is required in order to map numerical matrix entries to values of a color scale. After the normalization all values should lie in the interval [0,1]. Therefore each matrix entry was divided by the sum of all matrix entries. On the one hand this guarantees that all values will be positive and smaller or equal to 1, on the other hand in practice many values will be very small (much closer to 0 than to 1) after normalization. With the goal of weakening this effect we applied a transfer function to the normalized matrix which adapts to the data distribution in a much better way. In our case we did not linearly map the numerical data values to color values, but mapped the square roots of the numerical values to color values. For numbers n in the interval [0,1] it is always true that $0 \leq n \leq \sqrt{n} \leq 1$. This means the transfer function reserves a larger color range for the densely populated area of low values. The trade-off is a smaller differentiation in the sparsely populated area of high values. In addition, a bipolar color scale was chosen, ranging from bright yellow to dark blue. Bipolar color scales contain more distinguishable color shades than unipolar scales. An example for Finnish can be found in Figure 2. Here, the additional "+" symbols indicate that a vowel succession occurred more frequently than expected and the "−" symbols indicate the opposite case. This means that the matrix from Table 12 is integrated in the visualization.

FIGURE 2 The visualization represents the matrix from Table 8 but now in an automatically sorted manner. Two blocks of vowel transitions seem to appear frequently as the darker color in the upper left and lower right area suggests. In addition, it can easily be seen that /a/, /e/ and /i/ are the most frequent vowels.

In the matrix with the succession probabilities, in contrast, all values inherently lie in the interval [0,1] and thus do not have to be normalized. Nor is it necessary to apply a transfer function as the population of the number range is much less biased. See Figure 3 for an example. Again, "+" and "−" values are integrated into the visualization.

For the matrix showing the association strength (ϕ) values of vowel successions from Table 11 two unipolar color scales were used. Vowel successions occurring more frequently than expected were colored in blue and vowel successions that were less frequently observed than expected were given a red color. In this case the "+" and "−" symbols provide a redundant mapping. The higher the ϕ value was, the more saturated the color. Because of the skewed data distribution with many values close to 0, again a square root transfer function was applied. See Figure 4 for the Finnish example.

It has to be pointed out that a meaningful sorting of the matrix rows and columns is crucial for the visual analysis process. Only a proper matrix arrangement guarantees the visibility of phenomena like the blocks of vowels mentioned before. Figure 5 shows the direct visualization of the matrices from Table 9 and 11. This example shows that many interesting features are not clearly visible without sorting.

Interactive analysis

In the graphical user interface all matrices can be explored interactively. Zooming in and out allows more or fewer language matrices to be displayed on a screen. Detailed information on single vowel successions, like absolute values and probabilities, is displayed on demand by mouse-over, and so is the information which words from the corpus contain a certain vowel succession.

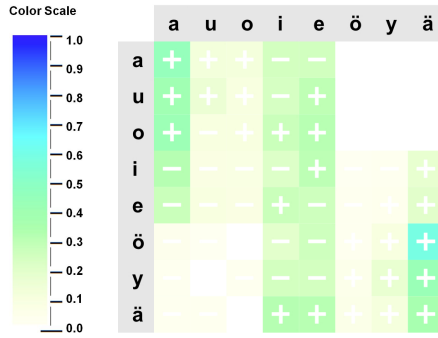

FIGURE 3 The visualization represents the matrix from Table 9 but now in an automatically sorted manner. Several interesting findings can be deduced from the visualization: (1) there are two blocks of vowels that almost never combine, viz. the block {a,u,o} and the block{ö,y,ä}; (2) after any vowel the vowels /i/ and /e/ are fairly probable to appear next; (3) after /a, u, o, e, i/ it is probable to observe an /a/ next. After /ä, y, ö, e, i/ it is probable to observe an /ä/ next. Especially the transition (ö-▷ä) is salient, it is very probable to observe an /ä/ after an /ö/. This effect is not visible in the matrix with the absolute frequencies (Figure 2), because /ö/ is a very infrequent vowel.

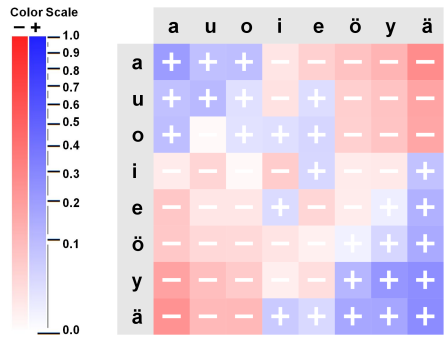

FIGURE 4 The visualization represents the matrix from Table 11 in combination with the matrix from Table 12. The latter determines the color hue to be displayed, blue for "+" and red for "−", with values of the first one affecting the color saturation: the higher the value the more saturated the color. Now, blocks of vowels that belong together can clearly be seen. As before, {a,u,o} build one block, {ö,y,ä} another independent block, and {i,e} cannot unambiguously be assigned to one of them.

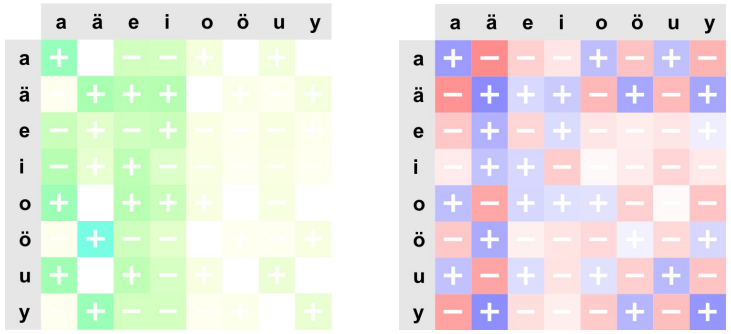

FIGURE 5 This visualization shows matrices that are left unsorted. They exactly correspond to the matrix from Table 9 (left) and to the matrix from Table 11 (right). The left matrix suggests that there are two groups of vowels which is misleading: the vowels on the left side are in general more frequent than the vowels on the right side, but they do not depend on each other in any way. The right visualization shows no easily perceivable pattern at all even though we know from Figure 4 that one exists.

5 Minimum amount of data necessary

For many less well-documented languages there are no large-scale textual resources like the Bible available. In order to find out whether our Visual Analytics approach can be applied to a wider range of underresourced languages, the scalability of the statistics and visualizations that we introduced were systematically tested on smaller text fragments. We investigated how many different words (types) are required so that we can derive reliable knowledge about VH in a language.

To do so, the resulting ϕ-matrices obtained by using a whole Bible text[7] were defined to be a gold standard for our purposes. The assumption is that for many languages this is an upper boundary on the size of corpora that one can get hold of. It does not mean that the results could not be refined with larger and larger textual resources, but the question is whether there is a certain point where the improvements converge and therefore justify the results for a language with this amount of data. We therefore tested for each language how closely the results obtained using smaller textual resources approached this gold standard. For each language we drew random type lists from the Bible, ranging from 10 up to 1,500 types — in intervals of 10 types.

In a next step, for each obtained type list a ϕ matrix was created in order to compare it to the previously created gold standard ϕ matrix. Then, for each amount of types — from 10 to 1,500 — the mean deviation of matrix entries

[7]Depending on the language under consideration the number of types ranges from 2,000 to 70,000.

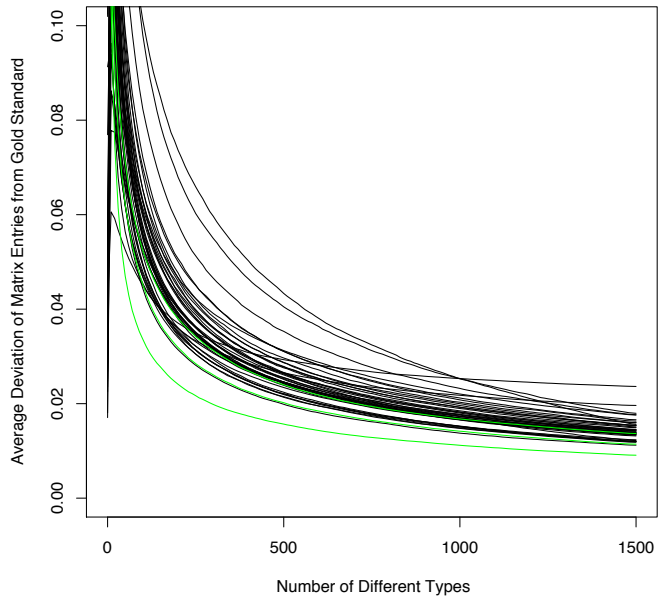

FIGURE 6 This plot shows the mean deviation of the ϕ matrix entries for smaller type lists from the gold standard entries (whole Bible type list). In order to smooth the curves and reduce the clutter we took the average curve of 1,000 trials. In total it shows the results for 42 languages, whereby the vowel-harmonic languages Turkish, Hungarian and Finnish are colored in green. Turkish has the quickest convergence among all languages.

from the gold standard was calculated. In Figure 6 this mean deviation is plotted on the y-axis against the number of types on the x-axis.

As can be seen in Figure 6, about 500 to 1,000 types are already sufficient to achieve a good convergence. The mean deviation of matrix entries (ϕ values) for almost all languages lies below 0.03 for an amount of 500 random types. The convergence depends on a set of different factors like the average word length, the number of vowels a language has and of course also the presence of VH. The well-known vowel harmonic languages Turkish, Hungarian and Finnish are among the languages that have a rather quick convergence (see the green lines in Figure 6). This confirms the assumption that in clear cases of VH languages, the harmonic pattern will already emerge from rather short texts. For languages with more marginal constraints on vowel distributions, such as those due to umlaut induced by certain suffixes in German, such

patterns will take more data to detect.

6 Wider cross-linguistic survey

In this section, some of our results for a larger number of languages are presented. Figure 7 shows the ϕ-matrices for some of the languages for which we a had a Bible corpus at our disposal. The languages are ordered according to the strengths of the effects they produce from left to right and top to bottom. Vowel harmonic languages tend to have rather saturated blue blocks along the diagonal (left top to right bottom) and rather saturated red blocks on the inverse diagonal. Most languages in the first two rows show this effect at least to a certain extent. Yet, from the second line onwards diagonal orientation is slightly less distinct and the color saturation gets less intense. Although there may be some effects visible in the further matrices, it is the first couple of matrices which are the most interesting ones. The languages were sorted from left to right and top to bottom according to an automatically determined value that indicates their tendency for harmony-like patterns, the average (absolute) ϕ value. A manual ordering following the visual salience would most probably have led to a very similar result.

A look at the first two probability matrices (see Figure 8) can also reveal interesting information. It shows that languages known to have strict VH (like Turkish, Finnish and Hungarian) are very prohibitive with respect to non-harmonic vowel successions. These simply do not occur and therefore large very bright areas appear in the upper right and lower left parts of the matrices. In Turkish the effect again is so strong that VH could be detected from the probability matrices only. The white patches in the probability matrices distinguish VH languages from those languages that have a superficially similar phenomenon (viz. partial reduplication), such as Maori.

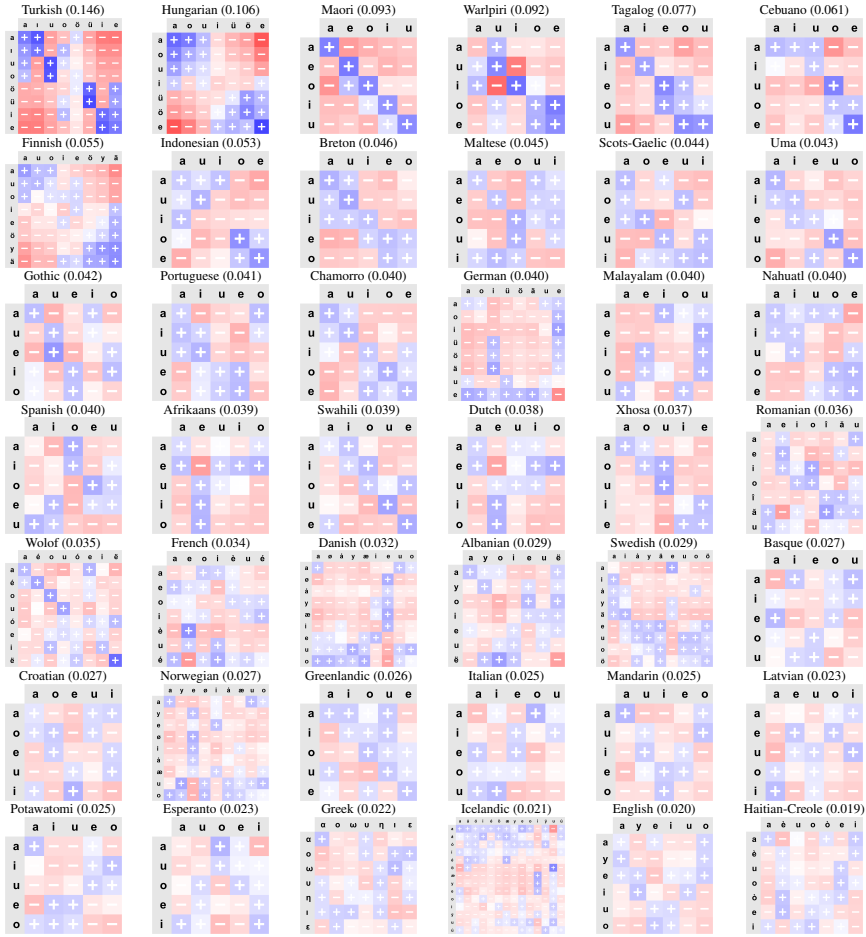

FIGURE 7 The ϕ matrices for 42 languages ordered according to decreasing average (absolute) ϕ values (rounded average in parentheses) from left to right and top to bottom.

6.1 Case study: Turkish and Finnish

The VH patterns for Turkish and Finnish were described in Sections 2.1 and 2.2, respectively. The ϕ-matrices that have been generated based on the Bible texts of the languages clearly visualize these patterns (see Figure 9).

The Turkish matrix shows the palatal harmony as two complementary blue blocks in the /a/- and /e/-columns whereas the labial harmony clusters are represented as adjacent 2-cell blocks in the /ı/-, /u/-, /ü/- and /i/-columns and also cover all rows of the matrix, indicating that there are no neutral vowels in the harmony processes present. The fact that the rows of the matrix can be filled twice with blue blocks shows that two harmony processes (labial and palatal) are active in the language.

The Finnish matrix shows a less clear-cut picture. Nevertheless two main blocks (in the upper left and bottom right corner) are visible and illustrate the harmony clusters for front (upper left) and back (bottom right) vowels. Unlike in Turkish, the harmony blocks are separated by two rows and columns in the middle of the matrix (representing the vowels /e/ and /i/), which indicates that the harmony involves neutral vowels. Recall that the matrix rows and columns have been sorted automatically and have not been arranged with the knowledge of which blocks should stand out.

6.2 Case study: Warlpiri VH vs. Maori reduplication

Turkish and Finnish are well-known VH languages, and their matrices clearly reveal the relevant patterns. Now, one of the motivations for this investigation was that it would be useful, in the absence of any prior information, to be alerted to languages showing harmonic or harmony-like patterns. For this purpose a variety of languages have been visualized with the same method (see Figure 7). For some languages that are not well known as having VH some striking patterns in the visualizations could be related to linguistic findings in the respective grammars.

Maori does not have VH, but an effect not altogether dissimilar is produced by its morphology of reduplication. Partial reduplication affixes an abstract CV syllable, and these segments are then specified through features of the corresponding CV segments of the base. In contradistinction to VH, the vowels of the reduplicand and the base are in fact identical, rather than only sharing the harmonic features; also, there are no patterns of dispreferred vowel sequences in the case of reduplication alongside the preferences. This difference can best be seen when comparing the probability matrices in Figure 8, where reduplication languages such as Maori and Tagalog do not exhibit such restrictions (in the form of white patches in the matrices) which are typical for genuine harmony languages.

In the ϕ-matrix of Warlpiri (see Figure 10) there is a conspicuous block that involves the letters /u/ and /i/. Both vowels are not likely to occur together

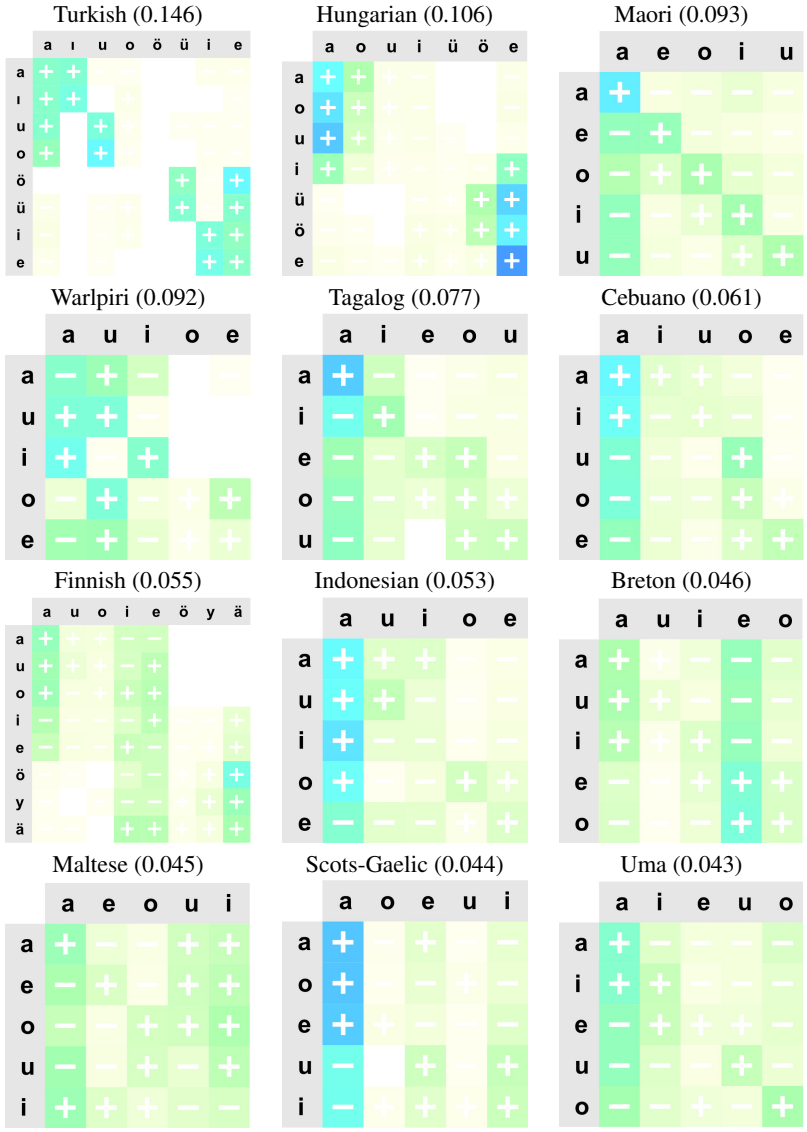

FIGURE 8 The probability matrices of the 12 top ranked languages. For languages known to have a strong harmony like Turkish, Finnish and Hungarian the matrices contain large very bright areas in the upper right and lower left. This is another visual characteristic that indicates VH because certain vowel successions are prohibited in vowel harmonic languages.

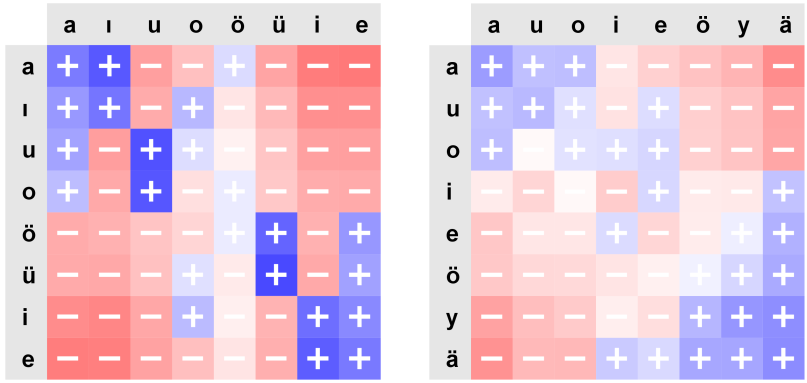

FIGURE 9 The ϕ-matrix for Turkish (left) and Finnish (right)

in words but rather have themselves as successor vowels. Indeed, Nash (1986) describes Warlpiri as having VH (both regressive and progressive) only involving the vowels /i/ and /u/.[8] Verbs with root-final /i/ change it to /u/ if the past tense suffix *-rnu* is added (regressive assimilation). Progressive assimilation changing /u/ to /i/ shows up with a large proportion of the nominal suffixes and enclitics. Consider the following words with the corresponding suffixes (see Nash 1986, p. 86):

1. kurdu-kurlu-rlu-lku-ju-lu
   child-PROP-ERG-then-me-they
2. minija-kurlu-rlu-lku-ju-lu
   cat-PROP-ERG-then-me-they
3. maliki-kirli-rli-lki-ji-li
   dog-PROP-ERG-then-me-they

As can be seen in 3., all subsequent suffixes change their vowels to /i/ if the last vowel of the stem is /i/.

Warlpiri therefore shows harmony in both directions. However, since our calculations do not take into account the direction of harmonic spreading, both harmony processes strengthen the results in Figure 10. The positive and negative cells in the matrix clearly show the non-co-occurrence of /i/ and /u/.

6.3 Case study: Udihe automatic VH detection

In this case study, we wanted to examine to what extent our approach is able to help researchers in detecting VH. We therefore chose to investigate Udihe,

[8] Notice that Warlpiri has a very small vowel inventory of only three vowels /a, i, u/. The occurrences of /e, o/ are due to loanwords or proper names from English. Notice that this difference is also reflected in the sorting of the matrix cells, where /a, i, u/ and /e, o/ are separate groups.

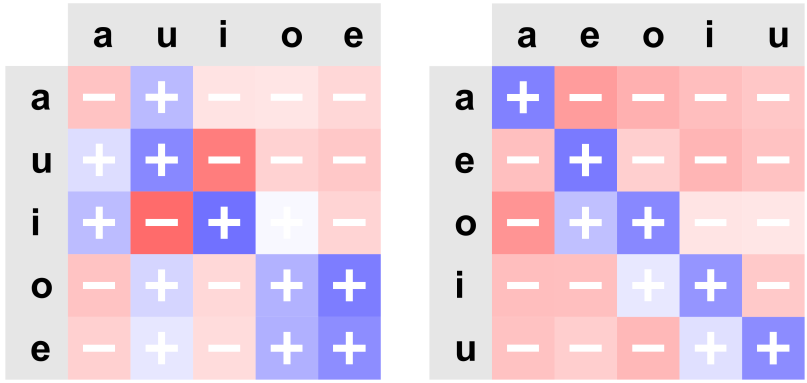

FIGURE 10 The ϕ-matrix for Warlpiri (left) and Maori (right)

another language from the Altaic family of languages (among them Turkish). Since Turkish and other Altaic languages have VH, there is a certain likelihood that Udihe will have VH as well. Yet we did not know beforehand what kind of harmonic patterns are active in the language. We obtained a text with a length of 2,450 words (1,200 types) which should be enough to detect reliable patterns according to the stability tests performed in Section 5.

In order to generate a hypothesis about possible vowel harmonic patterns, we must first ascertain whether harmony is present. We find three indicators for harmony:

- The average ϕ-value of Udihe (0.097) is the third highest after Turkish and Hungarian. This indicates that a strong effect like VH is present in the language.
- A look at the probability matrix (Figure 11) reveals that some successions are very probable and others very improbable — which is a characteristic of vowel harmonic languages.
- There are blue blocks along the diagonal as in Figure 11. The effect is not as clear as for the other vowel harmonic languages, though.

In Figure 11, we left out the vowel /ü/ because it appeared only in three successions within the corpus, so that no reliable statistics about it could be derived. As discussed above in Section 6.2, both the probability matrix and the ϕ-matrix are important for VH detection. If a vowel succession is very probable and at the same time has a highly positive association (ϕ value) this is an indication for a harmonic pattern. Clearly, this is the case for the transitions (a-▷a) and (ä-▷a) as well as (o-▷o) and (ö-▷o) as can be seen in Figure 11. As the vowel /i/ is very probable after any other vowel (except /ö/) it is unlikely to be a successor within a harmonic pattern. In both matrices

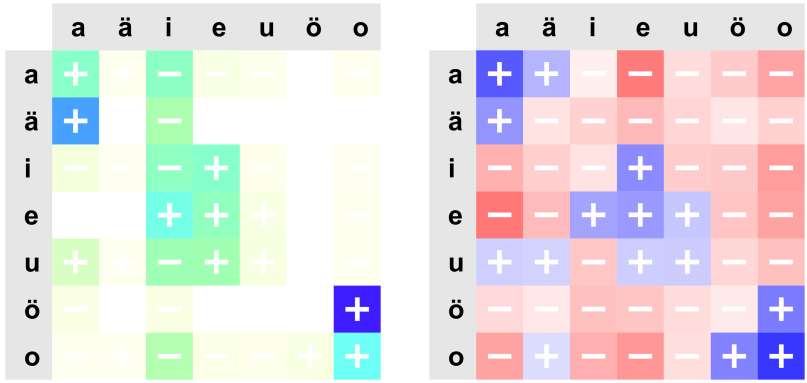

FIGURE 11 The probability matrix (left) and the ϕ-matrix (right) for the Udihe text fragment containing about 2,450 words (1,200 types).

the same block in the /e/ column is salient and indicates the harmonies (i-▷e), (e-▷e) and (u-▷e). The only further feature that is slightly conspicuous is the succession (u-▷a), but the effect is weaker than for (u-▷e). Table 15 summarizes these findings.

trigger		successor vowels
a, ä, (u)	→	a
o, ö	→	o
e, i, u	→	e

TABLE 15 Hypotheses about probable harmonies in Udihe

It has to be emphasized that the hypotheses shown in Table 15 were deduced without any prior knowledge about Udihe except that harmonic patterns could be expected there on grounds of family membership. The Udihe text was just fed into our program and the original text was not even looked at during hypothesis generation.

The accuracy of the results is very satisfying. Compared to the harmony processes of suffixes with respect to root vowels the predicted harmony patterns found in the visualizations (Table 15) correspond to what grammarians find in their analyses (Nikolaeva and Tolskaya, 2001, p. 74). This example shows that it is possible to generate accurate hypotheses about VH in languages within some minutes and without manually inspecting a single word in the language.

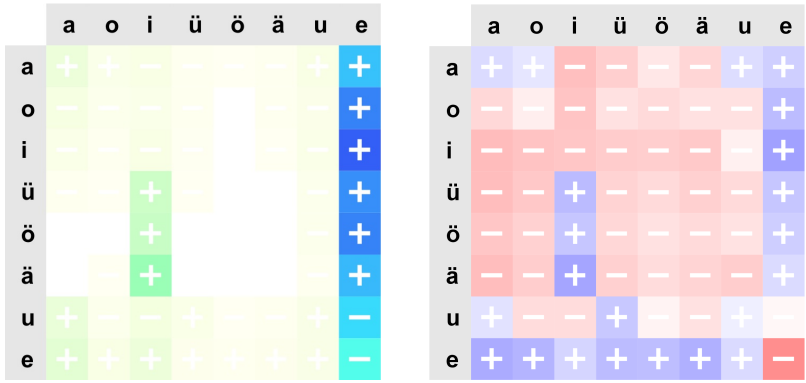

FIGURE 12 The probability matrix (left) and the ϕ-matrix (right) for German.

6.4 Case study: German umlaut

Umlaut in older Germanic was an assimilatory process whereby a back vowel changed to the associated front vowel when it was followed by an /i/ or /j/ (to only mention i/j-umlaut). Umlaut differs from the vowel harmony cases described so far in that only one vowel (the stressed vowel of the stem) is usually affected. Even though the harmonic process of umlaut is no longer active in the language and the former triggers for umlaut (/i/ and /j/ in the following syllable) have mostly disappeared in the relevant environments due to the weakening and loss of unstressed vowels, the general pattern is still visible even after the morphologization of umlaut, with umlaut now triggered by morphological categories. As can be seen in Figure 12, the umlauted vowels /ü, ö, ä/ occur more frequently before the vowel /i/ than before other vowels (except /e/, which is the most frequent successive vowel for all vowels). There are only a few suffixes left that still have an /i/ and trigger umlaut at the same time (e.g., *-in* as in *Französin* 'Frenchwoman' as compared to *Franzose* 'Frenchman', *-ig* as in *völlig* 'fully' as compared to *voll* 'full' or *-lich* as in *ärmlich* 'miserable' as compared to *arm* 'poor'). However, the relics of a once productive pattern of vowel assimilation can still be detected. Although German orthography does not quite mirror actual pronunciation, it appears to be sufficiently systematic to be indicative of such remnant syntagmatic dependencies among vowel distributions such as umlauting. Again, the German umlauting pattern can be detected entirely automatically, without manual inspection of actual German data.

7 Conclusion

The novel approach presented in this paper is an example for the fruitful integration of research from different disciplines. It shows that Visual Analytics is able to shed a sort of light on language data that will be found enlightening by linguists. In this respect the phenomenon of VH is a test case to see whether visualizations can provide a first analysis of data and can function as a useful tool to generate new linguistic hypotheses. Hence the goal of our visualizations is not to replace linguistic analysis but to support and guide it. We have demonstrated that a visualization of vowel patterns in terms of a quadratic succession matrix whose rows and columns are ordered automatically with respect to the similarity of vowels allows for an at-a-glance analysis of VH patterns across languages. This visual fingerprint of the language not only indicates if VH is present, but also which vowels are involved in the harmony processes and to what extent. The approach can easily be extended to investigate other dependencies (consonant co-occurrences etc.) within words.

The crucial characteristics of the visualizations produced here are (i) strongly blue colored blocks along the main diagonal of the ϕ-matrix and (ii) large white areas in the probability matrix, which indicate that some vowel successions that do not conform to harmonic patterns occur very infrequently. The latter characteristic is mostly used to distinguish VH languages from those having a similar type of vowel pattern where successions of identical vowels stick out due to (partial) reduplication processes where a syllable is copied for grammatical purposes and which differs from VH in not prohibiting certain vowel successions but only preferring some.

The visualization can be generated from simple plain text in a fully automatic manner. It is scalable in the sense that even a rather small number of types (about 500) is enough to derive linguistically meaningful results (Section 5). We have shown that our approach is detailed enough to allow for the visual identification of the precise harmony patterns involved (see Section 6.3 for the case study on Udihe) as well as for other related types of phenomena such as reduplication (see Section 6.2 on Maori) and German umlauting (Section 6.4).

Acknowledgments

This work has been funded by the research initiative “Computational Analysis of Linguistic Development” at the University of Konstanz and by the German Research Foundation (DFG) under the grant GK-1042, Explorative Analysis and Visualization of Large Information Spaces, Konstanz. The authors would like to thank Maria V. Tolskaya and Irina Nikolaeva for providing the Udihe texts and the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) for the Warlpiri Bible sections. We would also like

to thank Bernhard Wälchli and John Goldsmith for valuable comments on an earlier version of this paper. The first two authors are listed in alphabetical order: their contribution to this paper, the lion's share of working out the details, is equal.

References

Abbasi, Ahmed and Hsinchun Chen. 2007. Categorization and analysis of text in computer mediated communication archives using visualization. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '07)*, pages 11–18. New York, NY, USA: ACM.

Albrecht, J.S., R. Hwa, and G.E. Marai. 2009. The Chinese room: Visualization and interaction to understand and correct ambiguous machine translation. *Computer Graphics Forum (also in Proceedings of 2009 Eurographics/IEEE Symposium on Visualization,)* 28:1047–1054.

Altmann, Gabriel. 1986. Tendenzielle Vokalharmonie. *Glottometrika* 8:104–112.

Baker, Adam C. 2009. Two statistical approaches to finding vowel harmony. Tech. rep., University of Chicago.

Bertin, Jacques. 1983. *Semiology of graphics*. University of Wisconsin Press. ISBN 0299090604.

Clements, G. N. and Engin Sezer. 1982. Vowel and consonant disharmony in Turkish. In H. van der Hulst and N. Smith, eds., *The Structure of Phonological Representation*, chap. 2, pages 213–255. Dordrecht: Foris Publications.

Collins, Christopher, Sheelagh Carpendale, and Gerald Penn. 2009. Docuburst: Visualizing document content using language structure. In *Computer Graphics Forum (Proceedings of Eurographics, IEEE-VGTC Symposium on Visualization (EuroVis '09))*, vol. 28(3), pages 1039–1046.

Don, Anthony, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman, and Catherine Plaisant. 2007. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the sixteenth ACM Conference on Information and Knowledge Management (CIKM '07)*, pages 213–222. New York, NY, USA: ACM. ISBN 978-1-59593-803-9.

Fekete, Jean-Daniel and Nicole Dufournaud. 2000. Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In *Proceedings of the fifth ACM conference on Digital libraries (DL '00)*, pages 47–55. New York, NY, USA: ACM. ISBN 1-58113-231-X.

Gamon, Michael, Sumit Basu, Dmitriy Belenko, Danyel Fisher, Matthew Hurst, and Arnd Christian König. 2008. Blews: Using blogs to provide context for news articles. In *2nd AAAI Conference on Weblogs and Social Media*. American Association for Artificial Intelligence.

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

Goldsmith, John and Jason Riggle. 2007. Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. http://hum.uchicago.edu/ jagoldsm//Papers/boltzmann.pdf.

Goldsmith, John and Aris Xanthos. 2009. Learning phonological categories. *Language* 85(1):4–38.

Gregory, Michelle L., Deborah Payne, David McColgin, Nick Cramer, and Douglas Love. 2007. Visual analysis of weblog content. In *International Conference on Weblogs and Social Media. Published Online: http://icwsm.org/papers/3–Gregory-Payne-McColgin-Cramer-Love.pdf*.

Harrison, David, Emily Thomforde, and Michael O'Keefe. 2004. The vowel harmony calculator. http://www.swarthmore.edu/SocSci/harmony/public_html/.

Havre, Susan, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. 2002. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics* 8(1):9–20.

Hearst, Marti A. 1995. Tilebars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '95)*, pages 59–66. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co. ISBN 0-201-84705-1.

Henry, Nathalie and Jean-Daniel Fekete. 2006. Matrixexplorer: a dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics* 12(5):677–684.

Honkela, Timo, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. 1997. WEBSOM — self-organizing maps of document collections. In *Neurocomputing*, pages 101–117.

Honkela, Timo, V. Pulkki, and Teuvo Kohonen. 1995. Contextual relations of words in Grimm tales, analyzed by self-organizing map. In F. Fogelman-Soulie and P. Gallinari, eds., *Proceedings of International Conference on Artificial Neural Networks (ICANN-95)*, pages 3–7. Paris.

Keim, Daniel A., 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.

Keim, Daniel A. and Daniela Oelke. 2007. Literature fingerprinting: A new method for visual literary analysis. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology (VAST '07)*, pages 115–122. Washington, DC, USA: IEEE Computer Society. ISBN 978-1-4244-1659-2.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

Mayer, Thomas, Christian Rohrdantz, Frans Plank, Peter Bak, Miriam Butt, and Daniel A. Keim. 2010. 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.

Nash, David. 1986. *Topics in Warlpiri Grammar*. New York, London: Garland Publishing.

Nikolaeva, Irina and Maria V. Tolskaya. 2001. *A Grammar of Udihe*. Berlin: Mouton de Gruyter.

Oelke, Daniela, Ming Hao, Christian Rohrdantz, Daniel A. Keim, Umeshwar Dayal, Lars-Erik Haug, and Halldór Janetzko. 2009. Visual opinion analysis of customer feedback data. In *Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology (VAST '09)*.

Ringen, Catherine O. and Orvokki Heinämäki. 1999. Variation in Finnish vowel harmony: An OT account. *Natural Language and Linguistic Theory* 17:303–337.

Strobelt, Hendrik, Daniela Oelke, Christian Rohrdantz, Andreas Stoffel, Daniel A. Keim, and Oliver Deussen. 2009. Document cards: A top trumps visualization for documents. *IEEE Transactions on Visualization and Computer Graphics* 15(6):1145–1152.

Sukhotin, Boris V. 1962. Eksperimental'noe vydelenie klassov bukv s pomoščju evm. *Problemy strukturnoj lingvistiki* 234:189–206.

Thomas, James J. and Kristin A. Cook. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr. ISBN 0769523234.

Topping, Donald M. 1980. *Chamorro Reference Grammar*. Honolulu: The University Press of Hawaii.

van der Hulst, Harry and Jeroen van de Weijer. 1995. Vowel harmony. In J. Goldsmith, ed., *The Handbook of Phonological Theory*, chap. 14, pages 495–534. Oxford: Basil Blackwell Ltd.

Wanner, Franz, Christian Rohrdantz, Florian Mansmann, Daniela Oelke, and Daniel A. Keim. 2009. Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008. In *Proceedings of the IUI'09 Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009). Published Online: http://ceur-ws.org/Vol-443/paper7.pdf*.

Wattenberg, Martin and Fernanda B. Viégas. 2008. The Word Tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics* 14(6):1221–1228.

Wise, James A. 1999. The ecological approach to text visualization. *Journal of the American Society for Information Science* 50(13):1224–1233.

Wise, James A., James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. 1999. Visualizing the non-visual: spatial analysis and interaction with information for text documents. In *Readings in information visualization: using vision to think*, pages 442–450. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1-55860-533-9.

Wyszecki, Günther and W. S. Stiles. 2000. *Color Science: Concepts and Methods, Quantitative Data and Formulae (Wiley Series in Pure and Applied Optics)*. Wiley-Interscience, 2nd edn. ISBN 0471399183.