LEITFADEN VI Statistische Analyse in der Linguistic Landscape und Multimodalitätsforschuung

MANUAL VI: Statistic analysis in Linguistic Landscape and multimodal research

PD Dr. phil. habil. Ilona Schulze

Zweisprachiger Leitfaden über die statistische Auswertung von multimodalen Daten in der Linguistic Landscape Forschung

Bilingual manual on the statistical analysis of multimodal data in linguistic landscape research

August 2021

INHALTSVERZEICHNIS/CONTENTS

1 EINLEITUNG 3

2 SPLITTING UND LUMPING 4

3 DATENANALYSE 5

3.1 PROZENTUALER ANTEIL 8

3.2 TABELLEN 9

3.3 DIAGRAMME 9

3.4 MULTIVARIATE ANALYSEN 10

Explorative Verfahren 12

induktive Verfahren 17

1 INTRODUCTION 18

2 SPLITTING AND LUMPING 19

3 DATA ANALYSIS 20

3.1 PERCENTAGES 23

3.2 TABLES 24

3.1 DIAGRAMS 24

3.1 MULTIVARIATE ANALYSIS 25

Explorative Verfahren 27

statistical inference 32

4 ABBILDUNGEN/FIGURES 33

1 EINLEITUNG

Dieser Leitfaden erweitert das statistische Handwerkszeug zur Datenanalyse. Er geht eingangs kurz auf die unterschiedlichen Herangehensweisen in der Kategorisierung von Daten ein. Nach dem letzten Workshop scheint es dringlich geboten, hier noch einmal ausführlicher darauf einzugehen. Beide Verfahren und ihre Auswirkungen sollten wirklich jedem bewusst sein.

Im weiteren werden zunächst kurz die bekannten Verfahren noch einmal angesprochen und auf ihre Tauglichkeit in der multivariaten Datenanalyse hin betrachtet. Abschließend wird dann diese neu eingeführt und erläutert.

Sofern keine anderen Themenvorschläge kommen, werden diese Verfahren den nächsten Workshop prägen. Der Inhalt dieses Leitfadens wird dann als bekannt vorausgesetzt.

2 SPLITTING UND LUMPING

Splitten und Lumpen beschreibt zwei gegensätzliche wissenschaftliche Herangehensweisen an die Kategorisierung von Daten. Splitten leitet sich vom englischen Verb „to split“ und lumpen vom englischen „to lump“ ab, womit die zentralen Merkmale der beiden Verfahren bereits eindeutig benannt werden.

Beim Lumpen werden nur sehr zurückhaltend neue Kategorien eingeführt. Es wird zunächst versucht, alle Items in die bestehenden Kategorien einzugliedern. Nur wenn dies aufgrund großer Abweichungen nicht möglich ist, werden neue Kategorien gebildet. Dabei gibt es zunächst eine „Restekategorie“ wie „andere“ o.ä. Sollten sich jedoch neue Gruppen von Items bilden lassen, wird für diese dann eine neue eigene Kategorie geschaffen. Die Betonung liegt beim Lumpen also auf den Gemeinsamkeiten.

Anders verhält es sich beim Splitten. In diesem Verfahren werden schnell neue Kategorien gebildet, wenn Abweichungen festgestellt werden. Dies führt häufig zu einer hohen Zahl an Kategorien, die im Extremfall zur Einzelfallbeschreibung führen. Bei quantitativen Vorhaben kann dieses Verfahren kontraproduktiv sein.

Die Frage, wie fein Kategorien gegliedert werden und wie groß der Unterschied zwischen Kategorien sein muss, hängt einerseits von der Datenmenge ab und andererseits von der Forschungsfrage. Bei großen Datenmengen ist eine zu starke Feingliederung problematisch, weil sie die Gefahr birgt, den Überblick über die Gesamtdaten zu verlieren und keine allgemeingültigen Aussagen generieren zu können. Eine zu grobe Kategorisierung birgt hingegen die Gefahr, dass wichtige Informationen

verlorengehen. Daher ist es wesentlich, den Zuschnitt von Kategorien und die zulässigen Ausprägungen der Merkmale, die für die Kategorienzugehörigkeit unerlässlich sind genau zu definieren. Auch muss festgelegt werden, welche Merkmale für eine neue Kategorie sprechen und wie stark diese von den bereits bestehenden abweichen muss.

Besonders bei heterogenen Daten und Erhebungsräumen ist hier auch ein besonderes Augenmerk darauf zu legen, dass eine sich aus den Datentypen ergebende, unterschiedliche Anzahl von Kategorien nicht zusätzlich verschärft wird.

3 DATENANALYSE

Die Datenanalyse bewegt sich im Bereich der deskriptiven Statistik und spiegelt damit den empirischen Charakter des Projekts und der Daten wieder. Allgemein dient die deskriptive Statistik dazu, durch Tabellen, Grafiken oder Parametern umfangreiches Datenmaterial in nachvollziehbarer Art und Weise zu präsentieren. Gerade bei großen Datenmengen ist dies nicht nur für die Präsentation der Daten für Dritte wichtig, sondern auch für den Forscher selbst, da ein vollständiger Überblick über die Gesamtheit der Daten und mögliche Zusammenhänge ab einem gewissen Datenumfang nicht mehr gegeben ist. Somit ist die Aufbereitung der Daten mit Hilfe der deskriptiven Statistik auch die Basis für die Interpretation der Daten, in die zusätzlich weitere Informationen einfließen.

Für die Analyse und nachfolgende Interpretation der Daten werden sind insgesamt mehrere Analyseschritte notwendig. Zunächst In dieser Phase werden alle für die Fragestellung relevanten Variablen einzeln betrachtet und

absolute und prozentuale Werte ermittelt oder über andere Verfahren analysiert. Bei einer hohen Anzahl von Variablen kann dieser Schritt bereits sehr zeitaufwändig sein. Es ist aber von großer Wichtigkeit hier sorgfältig zu arbeiten, da die hier ermittelten Werte die Basis für alle weiteren Analyseschritte sind. Fehler, die an dieser Stelle passieren, werden also durch den gesamten Analysevorgang mitgeschleppt und können, sollten sie sich kumulieren, zu einer Verzerrung der Ergebnisse führen. Mit Kumulation ist hier nicht nur gemeint, dass für mehrere Variablen fehlerhafte Berechnungen vorgenommen werden, sondern dass in späteren Analyseschritten u.U. zwei oder mehr fehlerhafte Werte aus diesem Analyseschritt kombiniert und zueinander in Beziehung gesetzt werden.

Gerade bei einer Verbindung von Daten aus der Linguistic Landscape Forschung mit zusätzlichen Informationen aus der Multimodalitätsforschung ist besondere Sorgfalt geboten. Es ist immer zu Fragen, in welcher Reihenfolge Daten unterschiedlichen Typs miteinander verknüpft werden und auf welcher Analyseebene dies erfolgt. So macht es einen erheblichen Unterschied, ob ich eine Schriftfarbe mit einer Sprache verknüpfe oder umgekehrt. Im ersten Fall werden alle Sprachen untersucht, die in einer bestimmten Farbe in der Linguistic Landscape auftreten. Dies schließt nicht aus, dass die gleiche Sprache noch in einer anderen Farbe in der Lingustic Landscpae präsent ist – es werden also nicht alle Vorkommnisse der entsprechenden Sprachen erfasst, sondern nur ihr Vorkommen in der Verbindung mit der Schriftfarbe grün. Völlig andere Daten und Informationen werden generiert, wenn analysiert wird, mit welchen Farben eine bestimmte Sprache in der Linguistic Landscape auftritt.

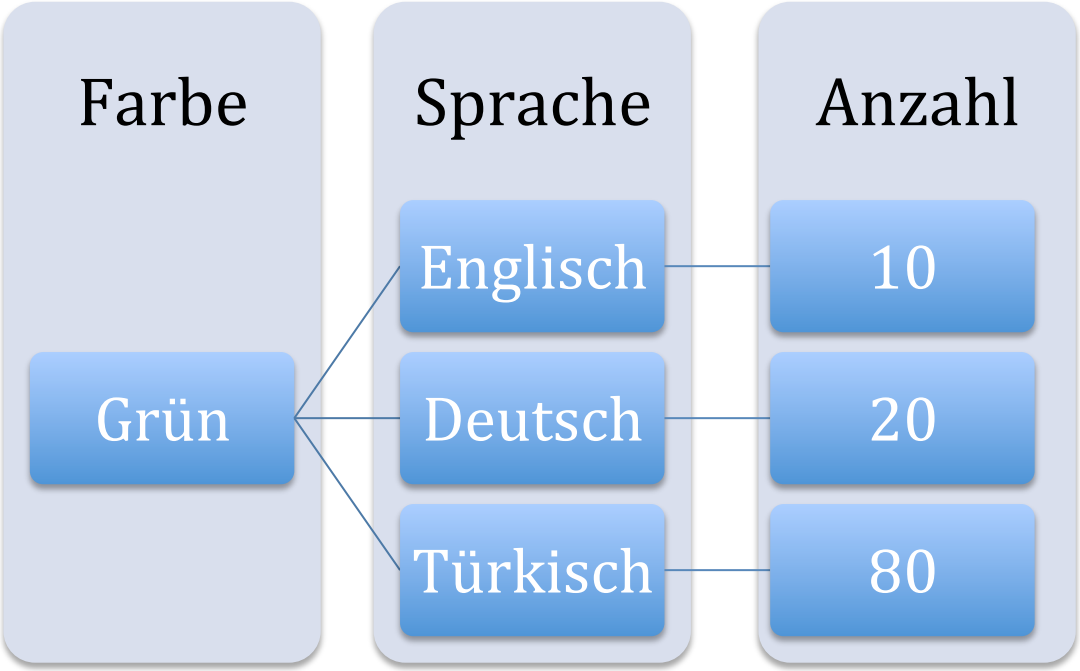

Abbildung 1: Fokus Farbe

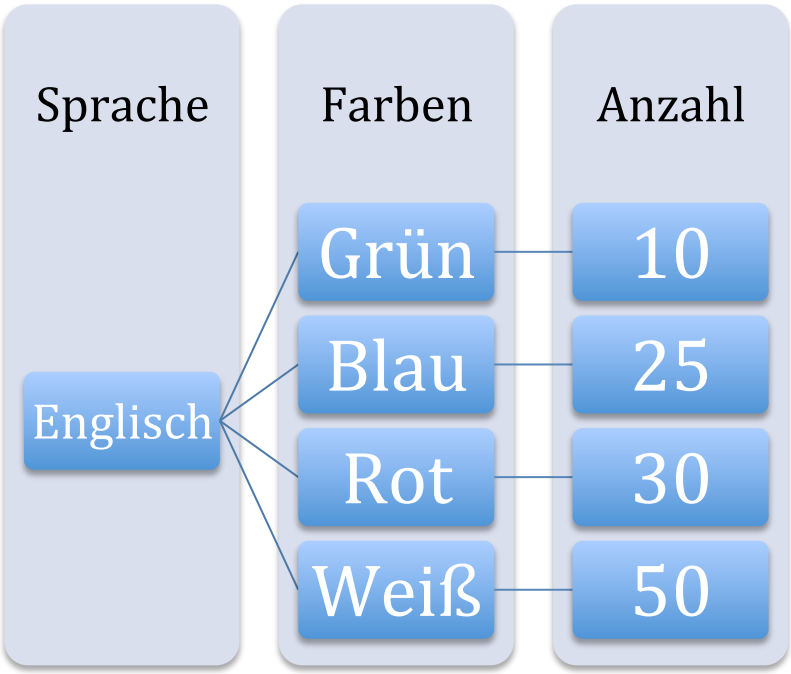

Abbildung 2: Fokus Sprache

3.1 Prozentualer Anteil

Streng genommen gehört diese Form der Darstellung nur am Rande zur deskriptiven Statistik, wird hier aber der Einfachheit halber dazu gezählt. Wie oben bereits erwähnt, können so innerhalb einer Variablen die Verhältnisse der einzelnen Ausprägungen zueinander beschrieben werden. Dies gilt auch für die Beschreibung von Beziehung zwischen unterschiedlichen Variablen, also Vorgehen, wie oben dargestellt. Für die Farbe Grün lässt sich festhalten, dass jeweils 10% auf das Deutsche und das Englische entfallen und 80% auf das Türkische. Für die Farben, in denen im obigen Beispiel Englisch auftritt ergeben sich von oben nach unten folgende Werte:

Grün: 8,7%

Blau: 21,7%

Rot: 26,1%

Weiß: 43,5%

Nicht mehr darstellen lässt sich die Hinzunahme einer weiteren Variablen. Prozentuale Darstellungen sind auf die Beschreibung der relativen Beziehung zwei Werte zueinander beschränkt. Was möglich ist, Ergebnisse als Zwischenergebnisse zu sehen und diese dann miteinander in Beziehung zu setzen. Es könnte also etwas über die Bedeutung der Farbe grün für die Darstellung englischsprachiger Elemente vermutet werden, wenn festgestellt wird, dass nur 10% aller Vorkommen von Grün auf Englisch sind und gleichzeitig von allen englischen Items nur 8,7 in grün sind.

3.2 Tabellen

Die Darstellung von Daten in Tabellen ist eine weitere Möglichkeit der Präsentation bzw. der Erfassung empirischer Daten. Eine Tabelle mit Zeilen und benannten Spalten wird *Matrix* genannt. In diese *Matrix* werden die Daten eingetragen. Dabei entsprechen die Zeilen in der Regel der Beobachtung und die Spalten einer Variablen der Daten. Im Grunde kann die Excel-Tabelle der Datenbankstruktur auch als *Matrix* bezeichnet werden. Hier wird dann auch das Problem von Tabellen deutlich: Selbst bei eher kleinen Datensätzen wird eine Tabelle schnell unübersichtlich und in den Daten vorhandene Strukturen oder Zusammenhänge können nicht oder nur sehr schwer erkannt werden. Gilt das Interesse nur einem Teil der Daten in der Tabelle kann ein Filtern der Daten helfen, aber auch hier gilt, dass es unübersichtlich wird, wenn die herausgefilterten Daten zu umfangreich sind.

Darüber hinaus gilt das für prozentuale Darstellungen gesagte: Die Berücksichtigung von mehr als zwei Variablen ist nicht möglich. Entsprechend ließen sich die obigen Beispiele zwar in einer Tabelle darstellen, und dies auch sowohl mit den absoluten als auch den relativen in einer Tabelle, aber die Hinzunahme einer dritten Variablen ist ausgeschlossen.

3.3 Diagramme

Diagramme können relative oder absolute Werte übersichtlich darstellen. Dazu müssen diese aber zusammengefasst werden, was in der Regel zu einem Informationsverlust führt. Allerdings bieten moderne Programme die Möglichkeit der dreidimensionalen Darstellung der Daten, so dass anders als bei den bereits besprochenen Formen, die Berücksichtigung einer dritten

Variablen möglich ist. Aber auch einfachere Abfragen können dreidimensional und damit übersichtlich dargestellt werden.

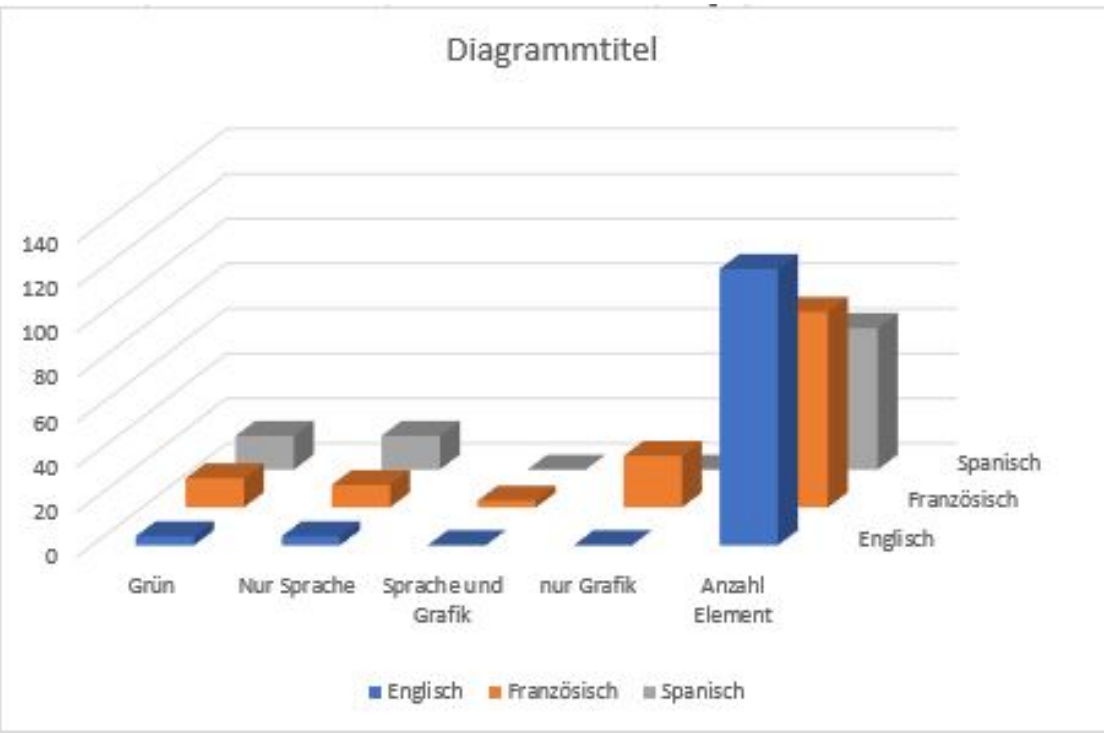

Abbildung 3: Diagramm dreidimensional

3.4 Multivariate Analysen

Einführende Informationen finden sich in Leitfaden V unter dem Punkt *Parameter*.

Multivariate Verfahren ermöglichen es, mehr als zwei Variablen zu untersuchen und sind damit das einzige Mittel, in dem dies direkt und ohne weitere Zwischenschritte erfolgen kann. Mit diesen Verfahren wird die deskriptive Statistik, wie sie im letzten Leitfaden beschrieben wurde, verlassen. Multivariate Analysen können Zusammenhänge und Abhängigkeiten zwischen Variablen aufdecken. Damit ist sie ideal für die Beschreibung komplexer

Zusammenhänge wie sie bei der gemeinsamen Betrachtung von Daten aus der Linguistic Landscape und der Multimodalitätsforschung gegeben sind, da anzunehmen ist, dass ein Phänomen mehrere Ursachen hat.

Im Wesentlichen sind hier Strukturentdeckende und strukturprüfende Verfahren. Zu den strukturentdeckenden Verfahren sind vor allem die Clusteranalysen die explorierende Faktorenanalyse (EFA = exploring factor analysis) und die Hauptkomponentenanalyse (PCA = Principal component analysis). Von den strukturprüfenden (induktiven) Analysen ist vor allem die multivariate Varianzanalyse (MANOVA) zu nennen. Die Darstellung erfolgt in der Regel über Streu- oder Netzdiagramme.

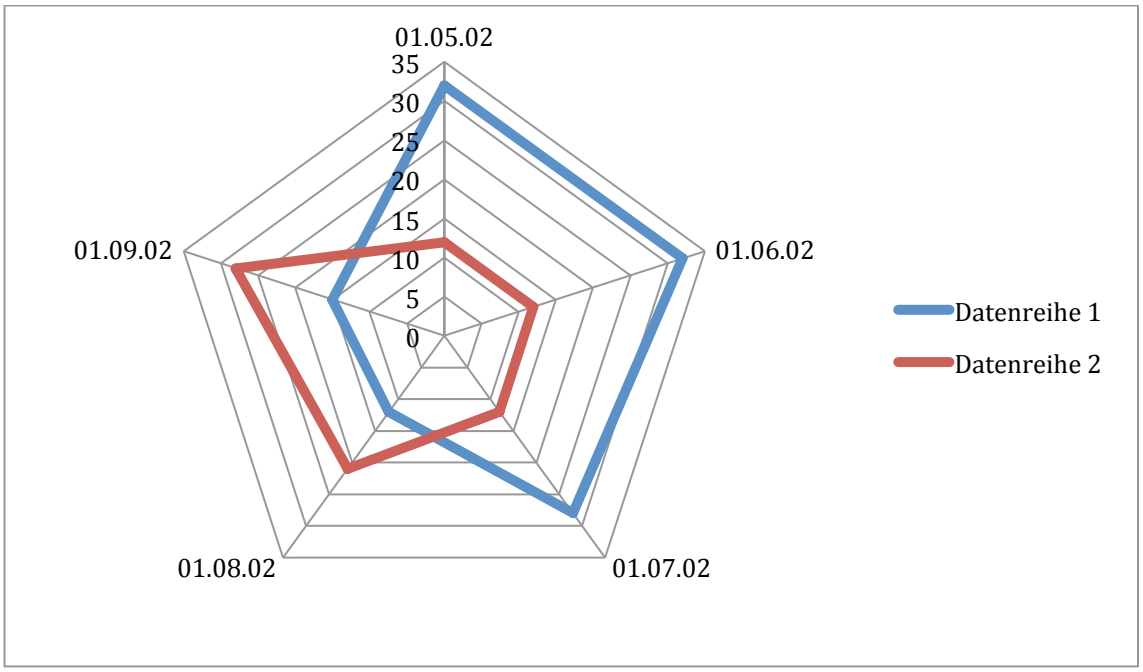

Abbildung 4: Beispiel Netzdiagramm.

EXPLORATIVE VERFAHREN

Strukturentdeckende Verfahren werden auch explorative Verfahren genannt. Wesentliches Merkmal explorativer Verfahren ist, dass sie nur mit geringen Vorannahmen arbeiten, also keine bis geringe Vermutungen über die Daten bzw. Beziehungen in den Daten zur Grundlage des Verfahrens machen. Dies ist insofern wichtig, als zu viele oder starke Vorannahmen über Beziehungen zwischen Variablen den Blick auf die Daten verstellen können, da diese Vorannahmen wie ein Priming wirken. Damit besteht die Gefahr, dass Zusammenhänge außerhalb des vermuteten Beziehungsgefüges nicht erkannt werden, da ein Priming eine wahrnehmungs- und handlungsleitende Wirkung entfaltet, also eine Voreinstellung gegenüber den Daten mit sich bringt. Sollen nun Strukturen aufgedeckt werden, ist es hinderlich, mit zu vielen Annahmen in die Analyse hineinzugehen; vielmehr sollte eine maximale Offenheit gegenüber den Daten und damit allen möglichen Beziehungen gegeben sein.

Explorative Verfahren sind damit in der Regel das Mittel der Wahl, wenn wenige Informationen über ein Forschungsfeld vorliegen, oder aber neuartige Verknüpfungen vorgenommen werden sollen. Hier kommt dann auch die Generierung von Hypothesen als besonderer Form der Annahme ins Spiel. Da im vorliegenden Projekt nicht nur eine komplexe Datenlage in Form von unterschiedlichen Arten von Daten gegeben sind, sondern auch bisher nicht oder nur äußerst selten berücksichtige Verknüpfungen dieser Daten in großem Umfang vorgenommen werden sollen, ist die Verwendung explorativer multivariater Verfahren in Betracht zu ziehen. Zwar werden diese in der Linguistic Landscape Forschung bisher nicht eingesetzt, was jedoch mit den in der Regel weniger komplexen Daten zu begründen ist.

EFA und PCA als die vielleicht relevantesten Verfahren führen häufig zu gleichen oder sehr ähnlichen Ergebnissen. Dies darf nicht darüber hinweg täuschen, das beiden Verfahren und Methode und Theorie grundlegnd unterscheiden.

Bei der EFA werden viele beobachtete Variablen auf einige wenige gemeinsame Faktoren zusammengeführt, die die Beobachtungen in einer Vielzahl von Variablen erklären. Dabei ist die EFA vorzugsweise für kardinalisierte Daten verwendet. Die PCA hingegen fasst die große Zahl an beobachteten Variablen zu so wenig Komponenten wie möglich zusammen und entspricht im Grunde einer Faktorenanalyse für Daten, die in Kategorien erfasst werden. So können Kontingenztabellen entweder mit grafischen Mitteln oder durch die Berechnung von Chi-Quadrat gemessen werden. Das Ergebnis sagt aus, ob die Datenverteilung zufällig ist, oder in einem Zusammenhang mit den Kategorien steht. In beiden Fällen spielen statistische (mathematische) Verfahren eine große Rolle. Dies genauer zu betrachten wird u.a. Gegenstand des nächsten Workshops sein.

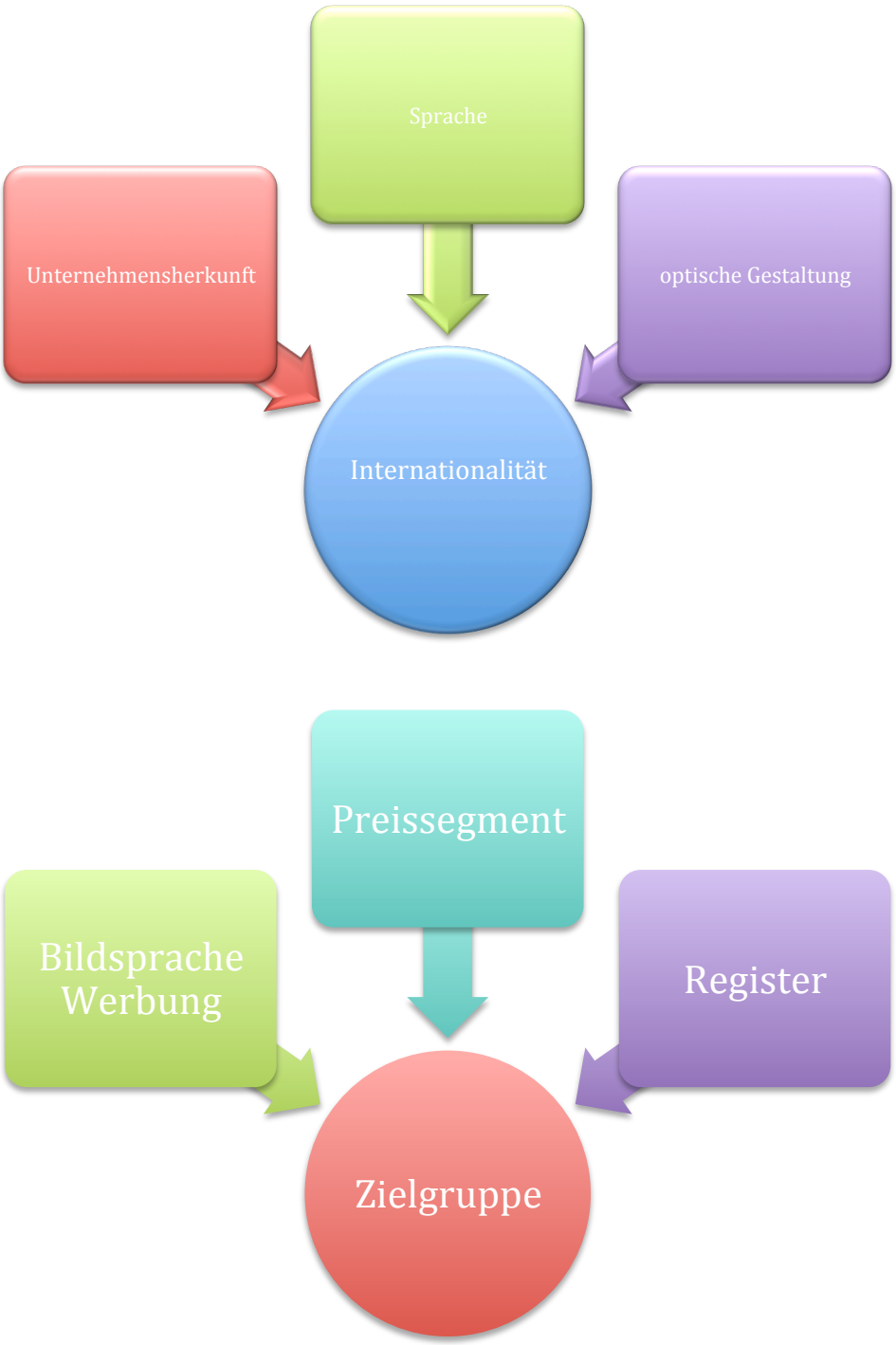

Abbildung 5: Beispiel PCA

Die Clusteranalyse setzt einen Schwerpunkt auf der Gruppierung von Daten. Dabei werden alle Fälle anhand vorgegebener Variablen in Gruppen eingeteilt,

wobei verschiedene Verfahren zur Verfügung stehen. Vereinfacht gesagt gibt es drei Formen der Gruppenbildung, die sich durch die Zuweisbarkeit von Elemente unterscheiden. Die erste Form hat feste Gruppen, die so zugeschnitten sind, dass jeder Fall nur einer Gruppe zugeordnet werden kann, die zweite Form setzt auf überlappende Gruppen, so dass jedes Elemente theoretisch mehr als einer Gruppe zugeordnet werden kann. Bei der dritten Form der Gruppenbildung (Fuzzy-Gruppen) wird jedes Element jeder Gruppe zugewiesen, aber mit unterschiedlichen Graden der Zugehörigkeit.

Die Zuordnung der Elemente zu den Gruppen ermöglicht eine Klassifizierung der Daten und macht Muster sichtbar. Die Darstellung der Ergebnisse erfolgt durch die Angabe von Ähnlichkeits- und Distanzmaßen auch eine grafische Darstellung ist möglich. Auch dieses Verfahren wird im nächsten Workshop genauer vorgestellt.

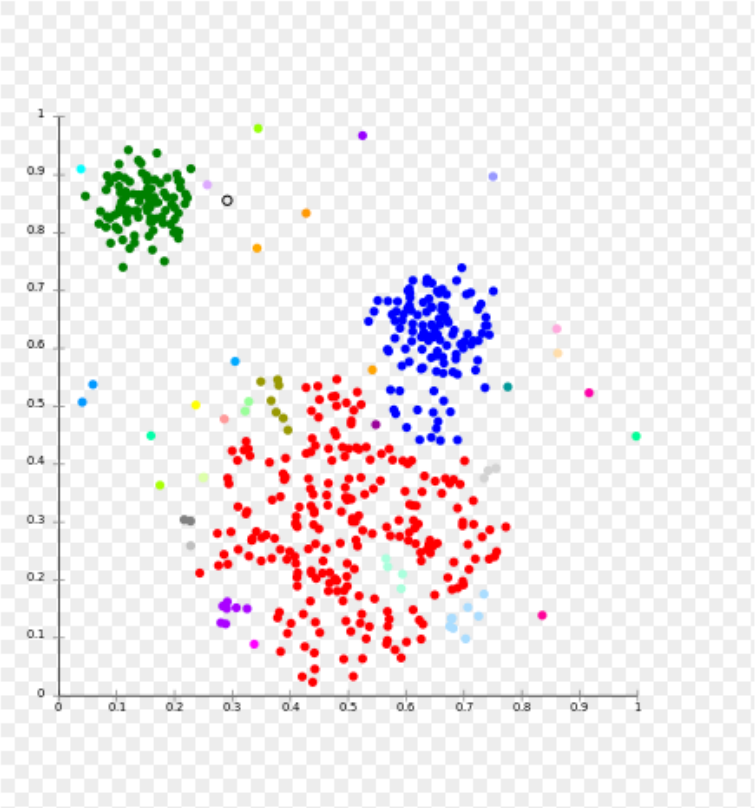

Abbildung 6: Beipiel Clusteranalyse Single Link-Verfahren.
https://creativecommons.org/licenses/by-sa/3.0/deed.en

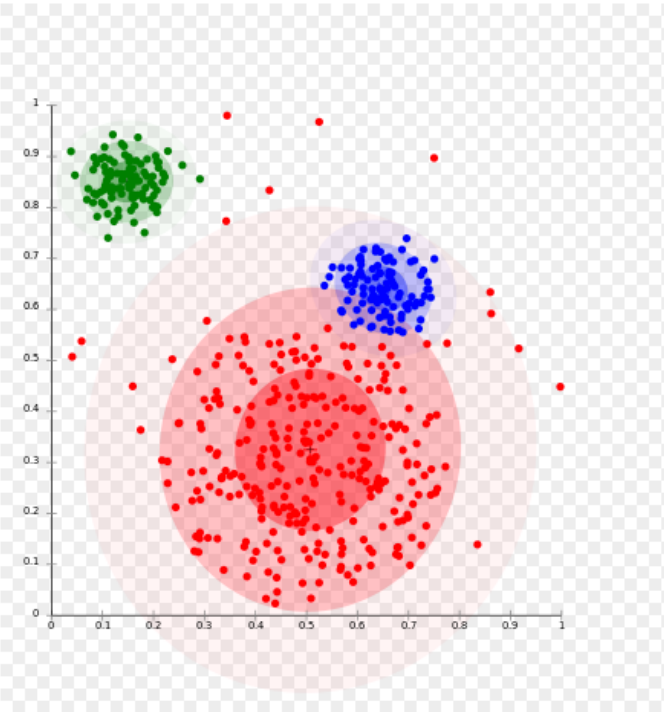

Abbildung 7: Beispiel Clusteranalyse Normalverteilung.
https://creativecommons.org/licenses/by-sa/3.0/deed.en

INDUKTIVE VERFAHREN

Mit der induktiven Statistik können Annahmen über das Gesamt der untersuchten Daten verifiziert oder falsifiziert werden; sie ist aber auch zur Modellentwicklung für Vorhersagen geeignet. MANOVA-Analysen untersuchen Unterschieden zwischen Gruppen auf mehreren abhängigen Variablen gleichzeitig, dabei sollten die Daten kardinalisiert sein. Hier könnte z.B. gleichzeitig überprüft werden, ob Markennamen und sonstige sprachliche Items jeweils die gleichen Präferenzen für Sprache, Farbgebung und grafischer Begleitung aufweisen. Für die Berechnung gibt es feste Formeln. Auch diese Verfahren werden Bestandteil des nächsten Workshops sein. Da sie aber nur eine untergeordnete Rolle für das Projekt spielen werden, werden sie nur der Vollständigkeit halber kurz angesprochen.

1 INTRODUCTION

This manual adds to the statistic tools for data analysis. It starts with briefly addressing the different approaches of categorizing data. This is a direct reaction on the discussion during the last workshop and its outcomes. It seems urgent to present this topic again. The knowledge of both approaches and their impact on data categorization are imperative.

After that the restrictions of the means of representation presented in the last manual are described with a focus on complex data. The following chapter than introduces multivariate data analysis.

In case no other topics are suggested for the next workshop in November, it will focus on multivariate data analysis. The content of this manual will be assumed as known.

2 SPLITTING AND LUMPING

The terms splitting and lumping describe two opposing scientific approaches for categorizing data. Splitting and lumping are derived from the respective verbs and very clearly describe the focus of both approaches

Lumping is very reluctant when it comes to introducing new categories. Before the introduction of a new category, any possibility of assigning an item to an existing category is checked. Only in case more than one item is not possible even when accepting major differences from standard elements, a new category is introduced. In the first place, single items are assigned to a category "other". If some of those items assigned to the category "other", can be grouped, they establish a new category. So, the focus of lumping is on focussing on commonalities.

The focus of splitting is just the opposite. This approach quickly establishes new categories in case of differences among items. This soon results in a high number of categories, which in extreme cases can cause the description if individual cases. This is counterproductive in quantitative research designs.

The amount of data and the research question decide how fine-grained categorization must be and which differences between the categories are acceptable. In case of big amounts of data too fined grained categorization is problematic as it bears the danger of loosing the overview of the total data so that no general assumptions about the data can be generated. In contrast, coarse categorization gives rise to the danger of important information getting lost. Thus, it is crucial to define precisely the layout of categories and the exact values for those parameters indispensable for category membership. Additionally, the parameters establishing new categories and their values have

to be determined as well as how strongly they have to differ from the existing categories and parameter values.

Especially in case of heterogeneous data and survey areas and the data they produce, special attention has to be paid to the danger of their complex categorization additionally increasing the number of categories.

3 DATA ANALYSIS

The methods of data analysis referred to here belong to the scope of descriptive statistics and mirror the empirical character of the project and the data collected. Generally speaking, descriptive statistics serves to represent vast quantities of data in a reproducible way with the help of tables, diagrams or parameters. This is not only important when presenting data to the scientific community. It is also a central aspect for the scientists themselves as it is virtually impossible to overlook big amounts of data in their totality. This bears the danger of missing relationships or patterns. Thus, using statistical methods to process empirical data also is a prerequisite for the interpretation of those data and combining them with additional information.

The analysis and subsequent interpretation of data requires several steps of analysis. In the first phase all relevant variables are examined individually, and the distribution of their values are given in absolute numbers and percentages or by means of any other appropriate method. This step can be time consuming if there are a large number of variables. Yet, it is important to carry out this step carefully as the results will be used in later steps. In the worst case, mistakes can cumulate and distort results. The accumulation of mistakes does not only mean, that the faulty results of more than one variable will have

in impact on the profile of an research area, but also that inaccurate results of two or more variables are combined in later steps and inevitably lead to results that do not reflect the data.

Special accuracy is necessary when connecting linguistic landscape data with additional information from multimodal research. It is of utmost importance to consider the influence the order in which data are connected has on the outcome. Additionally the level of analysis has to be taken into account, as the influence of order is conveyed to all following steps. It makes a significant difference whether the colour of writing is connected with a language or whether a language is connected to the colour of writing. In the first case, all languages appearing in a certain colour are the targets of such an analytical step. Hence not all findings of the languages encountered in the focus colour are of interest. In the second case the focus is on one language and the colours in which it appears in the public space. It is of no interest whether and if yes how many and which languages also are present in this language.

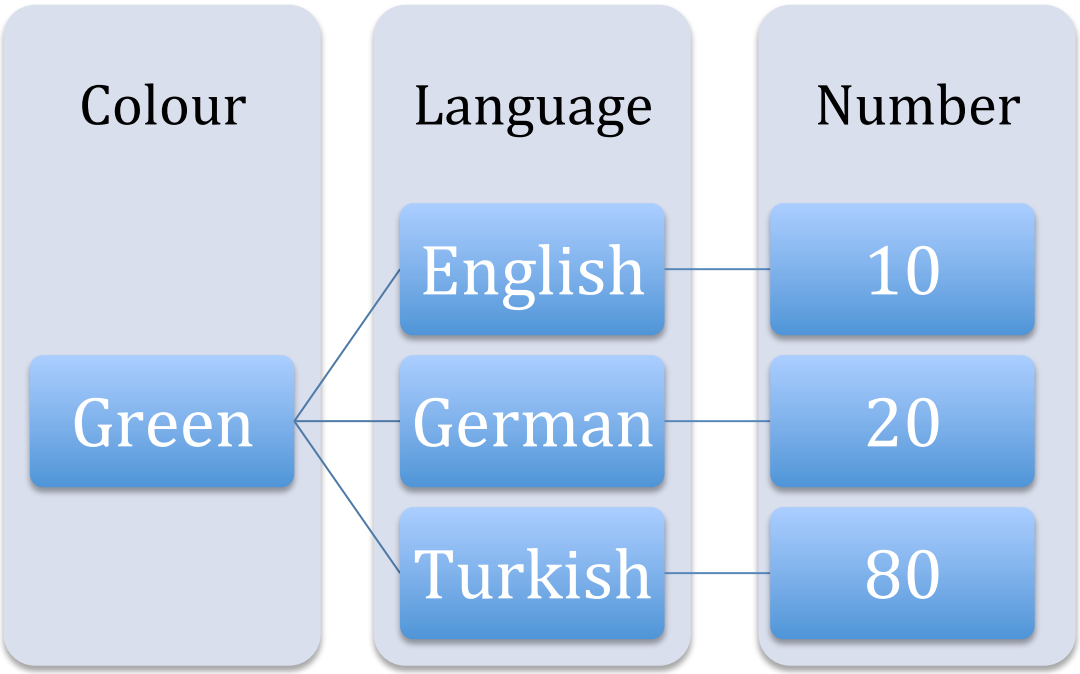

Abbildung 8: Focus on colour

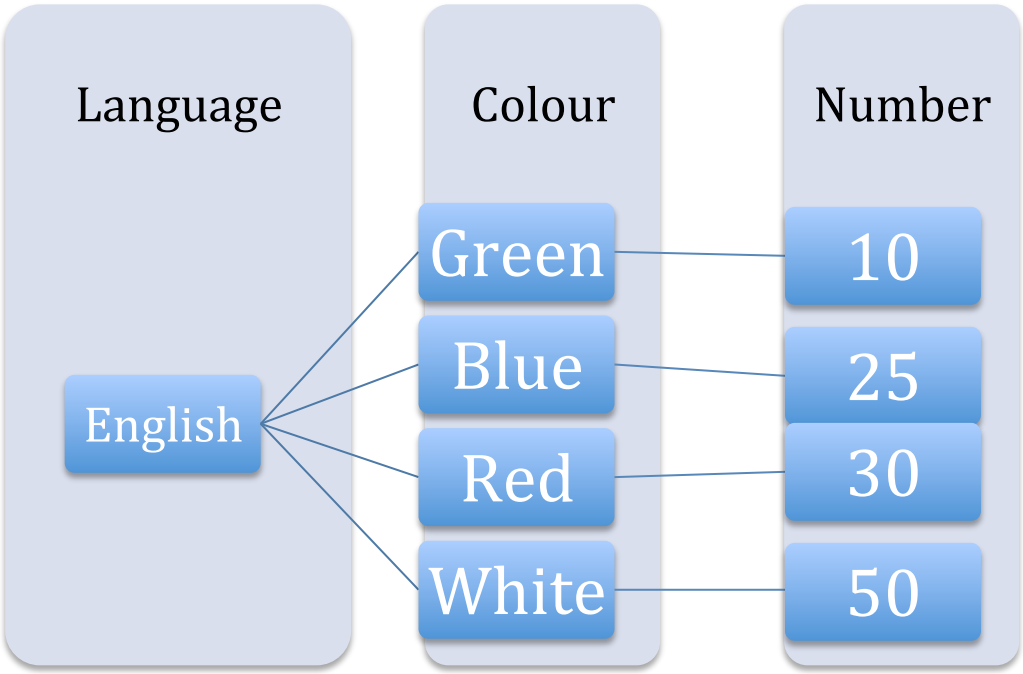

Abbildung 9: Focus on language

3.1 Percentages

Strictly speaking, this form of data representation is on the sidelines of descriptive statistics and is considered here for the sake of convenience. As mentioned above, percentages can be used to describe the shares the single values of a variable and thus make them comparable with each other.

The same holds for the description of relations between different variables, i.e. for approaches as those presented above. As for the colour green, English and German have a share of 10% each and Turkish has a share of 80%. The colours in which English is documented in the above example have the following shares:

Green: 8.7%

Blue: 21.7%

Red: 26.1%

White: 43.5%

It is not possible to add another variable. Indicating percentages is limited to the relative relation of the values. A frequently employed possibility is to split the analytical process in various steps and to calculate the percentages of every single step and then to interconnect those intermediate results. However, this is not always possible. So there some assumptions can be made about the meaning of the colour green for the presentation of English items, when data show that on the one hand only 10% green items are in English and that at the same time only 8.7% of all English items are green.

3.2 Tables

Showing data in a table is another form of entering empirical data. A table with rows and names columns is called *matrix*. Data are entered into this *matrix*. The rows in a *matrix* normally show the observation and the columns a variable. Basically, the Excel-table serving as a database to the project also can be called a *matrix*. Here the problem of tables becomes obvious: Even when dealing with small data sets, tables soon get confusing. Structures or relations get difficult to be observed or even go unnoticed. In case only a part of the overall data is of interest, tables can be filtered. Here too, the same holds: if the selected set of data is too big, it is difficult to manage.

In addition the remarks on percentage representation is true for tables, too. It is not possible to handle more than two variables. Accordingly, the above examples could be represented in a table, even with absolute an relative values in one table, but there is no way to include a third variable.

3.1 Diagrams

Diagrams are a means to clearly represent relative and absolute values. However, this requires merging them, which often results in a certain loss of information. However, modern programmes offer the possibility of three-dimensional presentation of data. Other than this is the case with the other forms of data representation presented so far, a third variable might be considered. But also simpler analysis can be represented in a three-dimensional way.

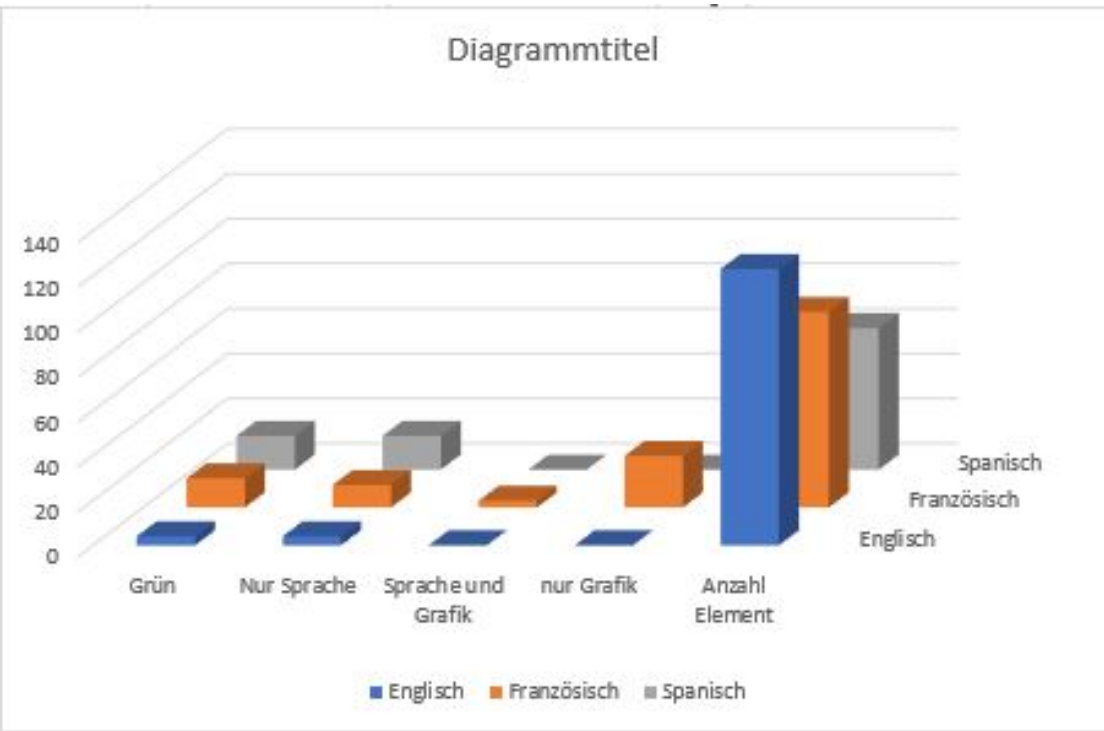

Abbildung 10: 3-dimensional diagram

3.1 Multivariate Analysis

Introductory remarks can be found in manual V in the section "Parameters". Multivariate approaches analyse more than two variables and thus are the only means to do so right away without any intermediate steps. Turning towards these approaches means leaving descriptive statistics as described in the last manual. Multivariate analysis can reveal connections and dependencies between variables. It is hence ideal for describing complex relationships such as they occur in the joint survey of data from linguistic landscape and multimodal research. Corresponding research questions often are based on the assumption that a phenomenon has more than one source.

Essentially, it is exploratory analysis and statistical inference that have to be mentioned. Exploratory approaches include cluster analysis, exploratory factor analysis (EFA) and principal component analysis (PCA). Furthermore, multivariate analysis of variance (MANOVA) has to be mentioned as a non-exploratory approach. Representation often is realized by means of scatter plots or radar charts.

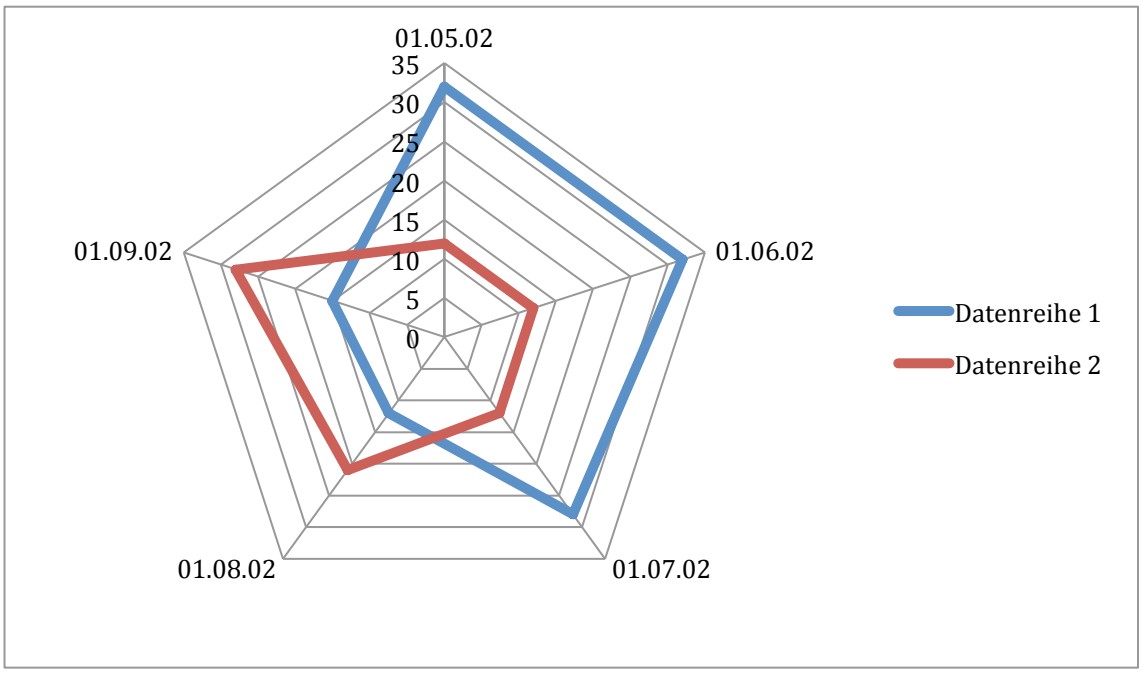

Abbildung 11: Example radar chart.

EXPLORATIVE VERFAHREN

A low number of assumptions about the data to be analysed is a central feature of exploratory approaches. They do no rely on presumptions about data or the relationships that might be found in the data. This is important as too many or too strong presumptions about the relations between variables can inhibit data analysis as they act like a sort of priming. Priming carries the danger of not detecting relations outside the assumed network as it influences perception and action and thus affects the attitude towards the data to be analysed. So, if the aim is to detect structures, too many presumptions are rather an obstacle. Rather, there should be a maximum of openness towards the data and thus any relationship.

Exploratory approaches are the first choice when there is only little information about a field of research or if new relationships are searched for. In this case generating hypothesis as a special form of presumptions is of importance. The present project does not only deal with complex data structures in form of different types of data but also aims at merging them in a new way and on a large scale. Against this background the use of exploratory approaches has to be taken into account. The fact that linguistic landscape research still ignores those approaches is due to the often less complex research questions that do not require "sophisticated" statistical analysis.

EFA and PCA are probably the most relevant approaches and often yield the same or very similar results despite their very different methods and theories.

EFA reduces a high number of observed variables to a restricted number of shared factors, which can explain the observations in the variables. PCA on the other hand merges the large number of observed variables to as few components as possible and resembles factor analysis. This latter approach groups data in categories. PCA uses contingency tables, which can be visualized with graphic means or by calculations like Chi-square. The result shows whether the distribution is random or if there is a correlation with the categories used. In both cases statistical (mathematical) approaches are of great importance. To take a closer look at those approaches will be part of the next workshop.

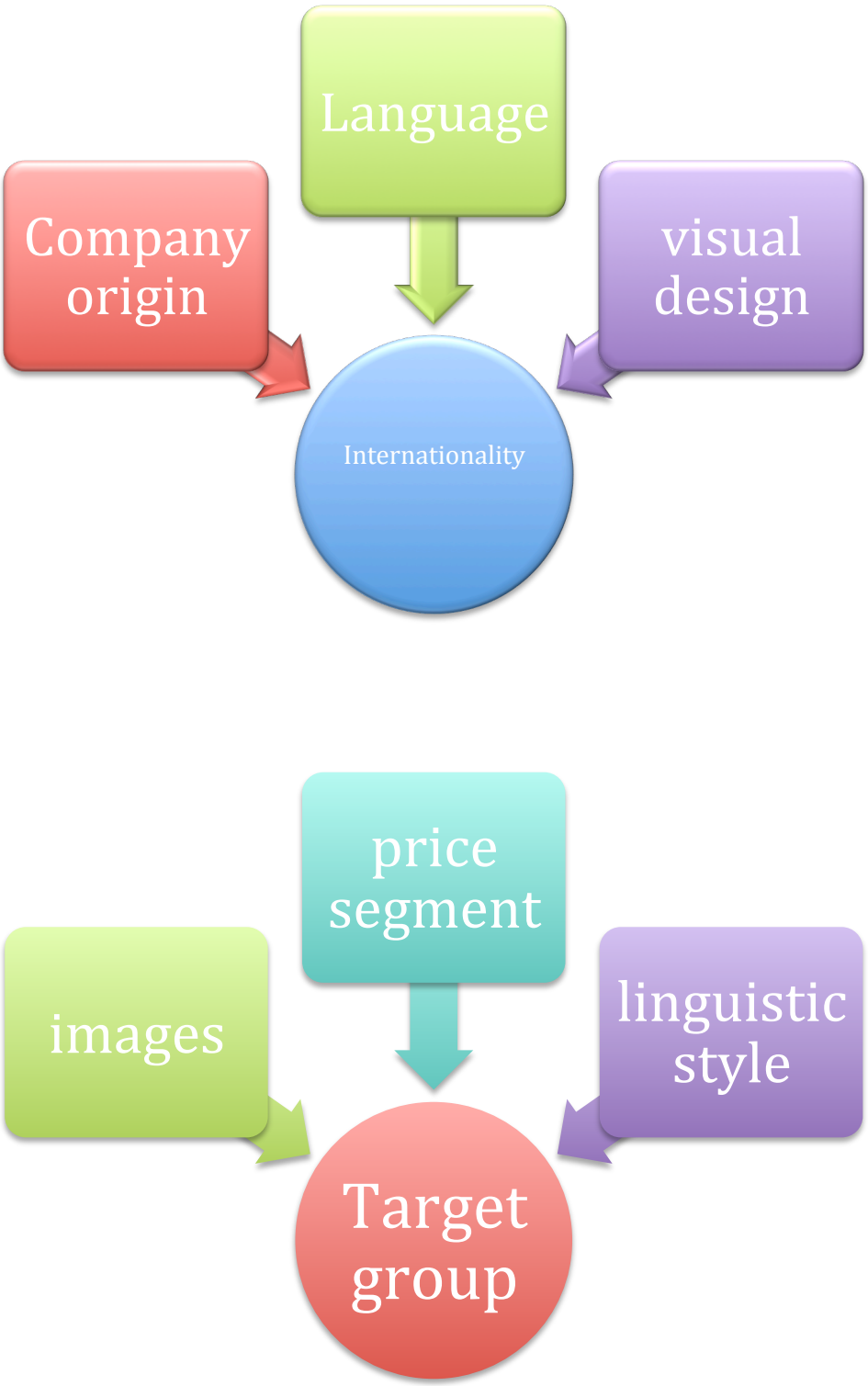

Abbildung 12: Example PCA

Cluster analysis focuses on grouping data. All cases are assigned to groups with the help of given variables. To put it in a simple way: there are three different ways for grouping, which differ in the way elements are assigned to a group. The first form works with fixed groups. They are established in a way that guarantees that each case can only be assigned to one group. The second form has overlapping groups. Hence, theoretically each element can be assigned to more than one group. The third form has groups that are called fuzzy groups. Each element is assigned to all groups but with different degrees of membership.

Assigning elements to groups allows to classify data and to detect patterns. The results are presented in form of similarity and/or distance measure. Graphic presentations are possible, too. The approaches mentioned will be part of the next workshop.

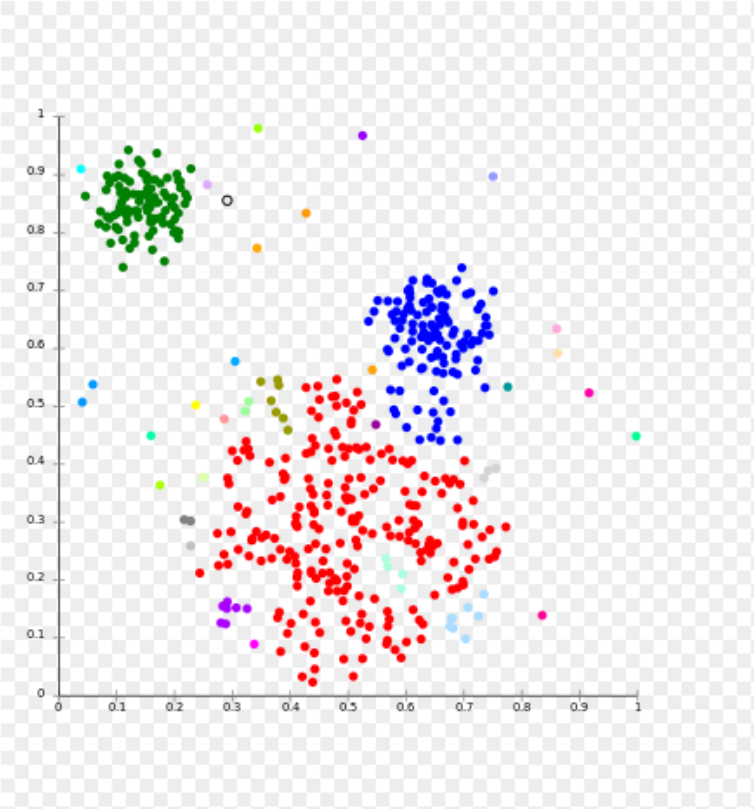

Abbildung 13: Example Clusteranalysis Single Linkage.
https://creativecommons.org/licenses/by-sa/3.0/deed.en

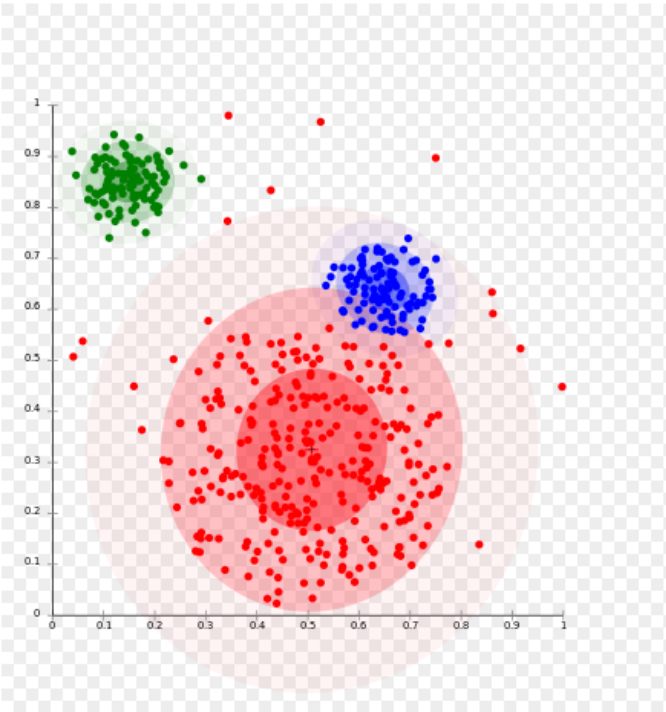

Abbildung 14: Beispiel Clusteranalysis normal distribution.
https://creativecommons.org/licenses/by-sa/3.0/deed.en

STATISTICAL INFERENCE

Statistical inference allows verifying or falsifying assumptions about the entire set of data. Additionally it is suited to develop prediction models. MANOVA-analysis examines the difference between groups on several dependent variables at the same time. This offers the chance to examine e.g. whether brand names and other linguistic items show the same preferences for language, colour and graphic components. Calculation is carried out with statistical formulae. This approach, too, will be part of the next workshop. As it is of minor importance for the project it will only be touched briefly and for the sake of completeness.

4 ABBILDUNGEN/FIGURES

Abbildung 1: Fokus Farbe 7

Abbildung 2: Fokus Sprache 7

Abbildung 3: Diagramm dreidimensional 10

Abbildung 4: Beispiel Netzdiagramm. 11

Abbildung 5: Beispiel PCA 14

Abbildung 6: Beipiel Clusteranalyse Single Link-Verfahren.
https://creativecommons.org/licenses/by-sa/3.0/deed.en 16

Abbildung 7: Beispiel Clusteranalyse Normalverteilung.
https://creativecommons.org/licenses/by-sa/3.0/deed.en 16

Abbildung 8: Focus on colour 22

Abbildung 9: Focus on language 22

Abbildung 10: 3-dimensional diagram 25

Abbildung 11: Example radar chart. 26

Abbildung 12: Example PCA 29

Abbildung 13: Example Clusteranalysis Single Linkage.
https://creativecommons.org/licenses/by-sa/3.0/deed.en 31

Abbildung 14: Beispiel Clusteranalysis normal distribution.
https://creativecommons.org/licenses/by-sa/3.0/deed.en 31