

# Deskription Survey Qualitaet

## Contents

<b>1</b>	<b>Einführung</b>	<b>1</b>
<b>2</b>	<b>Deskription</b>	<b>2</b>
2.1	Gesamter Datensatz . . . . .	2
2.2	Nominal -/ Ordinalskalierte Einflussgrößen . . . . .	2
2.3	Filter . . . . .	2
2.4	Metrisch Skalierte Einflussgrößen . . . . .	21
2.5	Outcome: Qualität . . . . .	22
<b>3</b>	<b>Modelle</b>	<b>23</b>
3.1	Genestete Variablen . . . . .	24
3.2	Modelle . . . . .	27

## 1 Einführung

Outcome des Projekts *Surveyqualität* ist die Qualität von Fragebögen, welche sich aus der Reliabilität und Validität ( $q^2 = r^2 * v^2$ ) zusammensetzt. Diese wurden durch ein Strukturgleichungsmodell *Multitrait-Multimethod* (MTMM) berechnet, vor allem mithilfe von großen Befragungen des *European Social Surveys* (ESS).

Die Qualität, sowie die Reliabilität und Validität sind hierbei stetige Merkmale, welche im Bereich [0; 1] liegen.

Um die Qualität von neuen Surveyfragen zu prognostizieren wurden über die Zeit mehrere Modelle berechnet, *Survey Quality Predictor* (SQP). Die erste Version dieser Vorhersage wurde mittels linearer Regression berechnet. In der zweiten Version mittels *random forest*. Die dritte Version wiederum mittels *random forest* (Schweisthal).

Um bei der Erstellung neuer Fragebögen den Forschern unter die Arme zu greifen, soll nun ein Regressionsmodell berechnet werden, damit Aussagen wie “Falls Sie eine Einleitung zu Ihrer Frage hinzufügen erhöht sich die Qualität um xy”. Also ein interpretierbares Modell.

## 2 Deskription

Im folgenden ein kurzer Überblick über alle Parameter, welche sich im Datensatz befinden und relevant zur Berechnung sind.

### 2.1 Gesamter Datensatz

Der Datensatz besteht aus **6074 Beobachtungen** (mit NAs: 8070), **60** zu verwendende **Variablen** (Outcome + Einfluss), wobei **42** Einflussvariablen **Nominal- und Ordinal** und **16** Einflussvariablen **Metrisch** skaliert sind.

Table 1: Variablen im gesamten Datensatz nach Skalenniveau

Datensatz	Anzahl
Gesamt	60
Nominal- / Ordinal	42
Metrisch	16

### 2.2 Nominal -/ Ordinalskallierte Einflussgrößen

Von den 42 Nominal -/ Ordinalskallierten Einflussgrößen sind die meisten **binär** kodiert (**29**), fast alle mit **weniger als 10 Kategorien** (**40**) und **2** mit **mehr als 10 Kategorien**.

Table 2: Häufigkeit von Nominal -/ Ordinalskallierten Variablen im Datensatz

Anzahl an Kategorien	Häufigkeit im Datensatz
2	29
3	5
4	3
5	1
6	2
11	1
29	1

Der Einfluss der Spalte mit **29** Kategorien (**Sprache**) ist von großer Wichtigkeit, da dieser verwendet werden soll, um **random intercepts** zu implementieren (hierarchische Struktur soll beachtet werden: Studien genestet in Experimenten in Ländern / Sprache).

Eine kleine Übersicht über die Daten können die folgenden Histogramme geben.

#### 2.2.1 Binäre Einflüsse - Unabhängig von Filtern

#### 2.2.2 Nicht binäre Einflüsse - Unabhängig von Filtern

### 2.3 Filter

Im folgenden Abschnitt gehe ich etwas näher auf die Filter ein. Die Grafiken sind aus *Codebook Routing* entstanden, in dem beschrieben wird, in was für einer Reihenfolge einzelne Fragebögen bewertet werden sollten.

### Binäre Einflüsse – unabhängig von Filtern

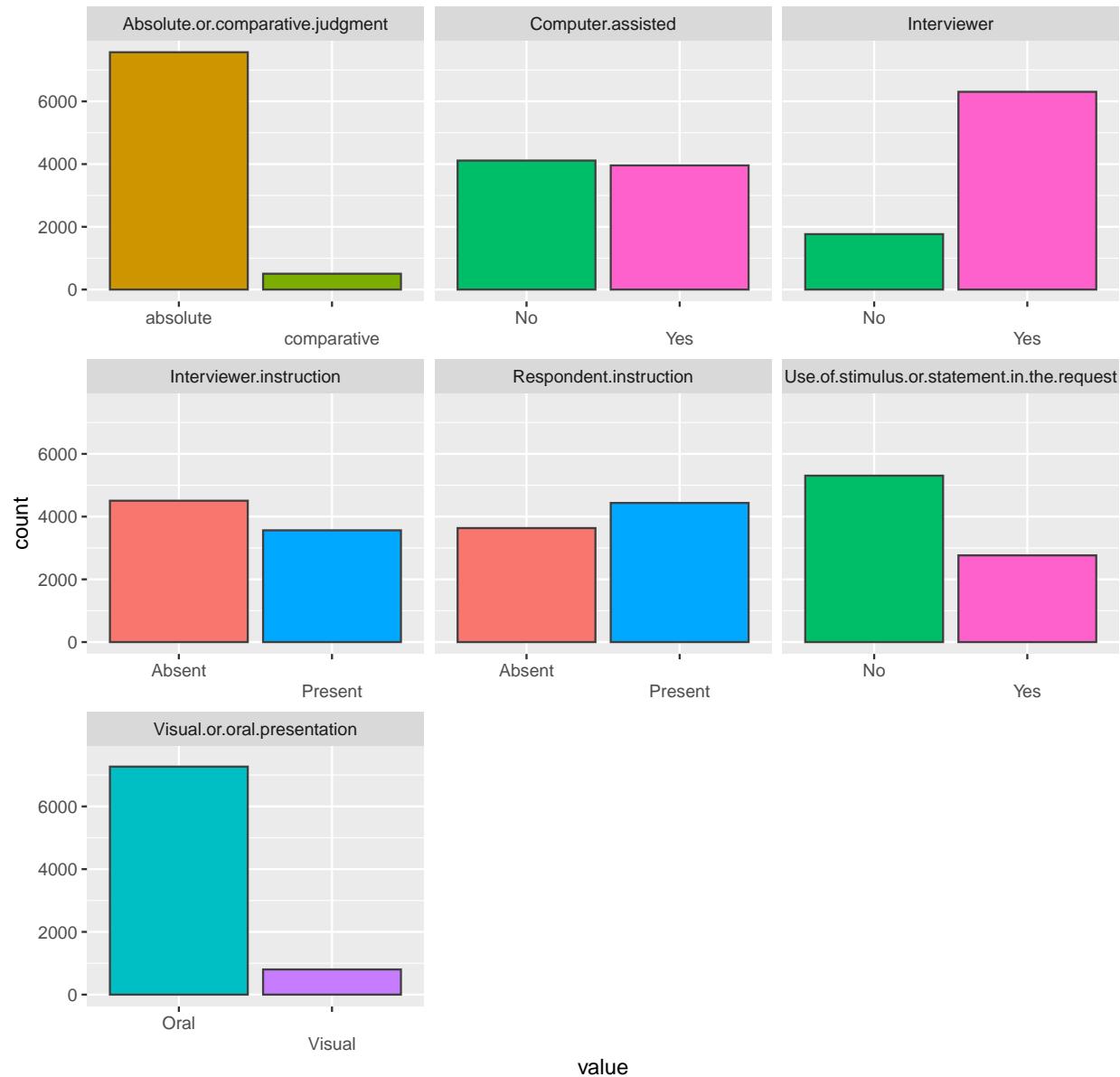


Figure 1: Binäre Einflussgrößen, welche nicht in Filtern vorkommen

## Nicht binäre Einflüsse – unabhängig von Filtern

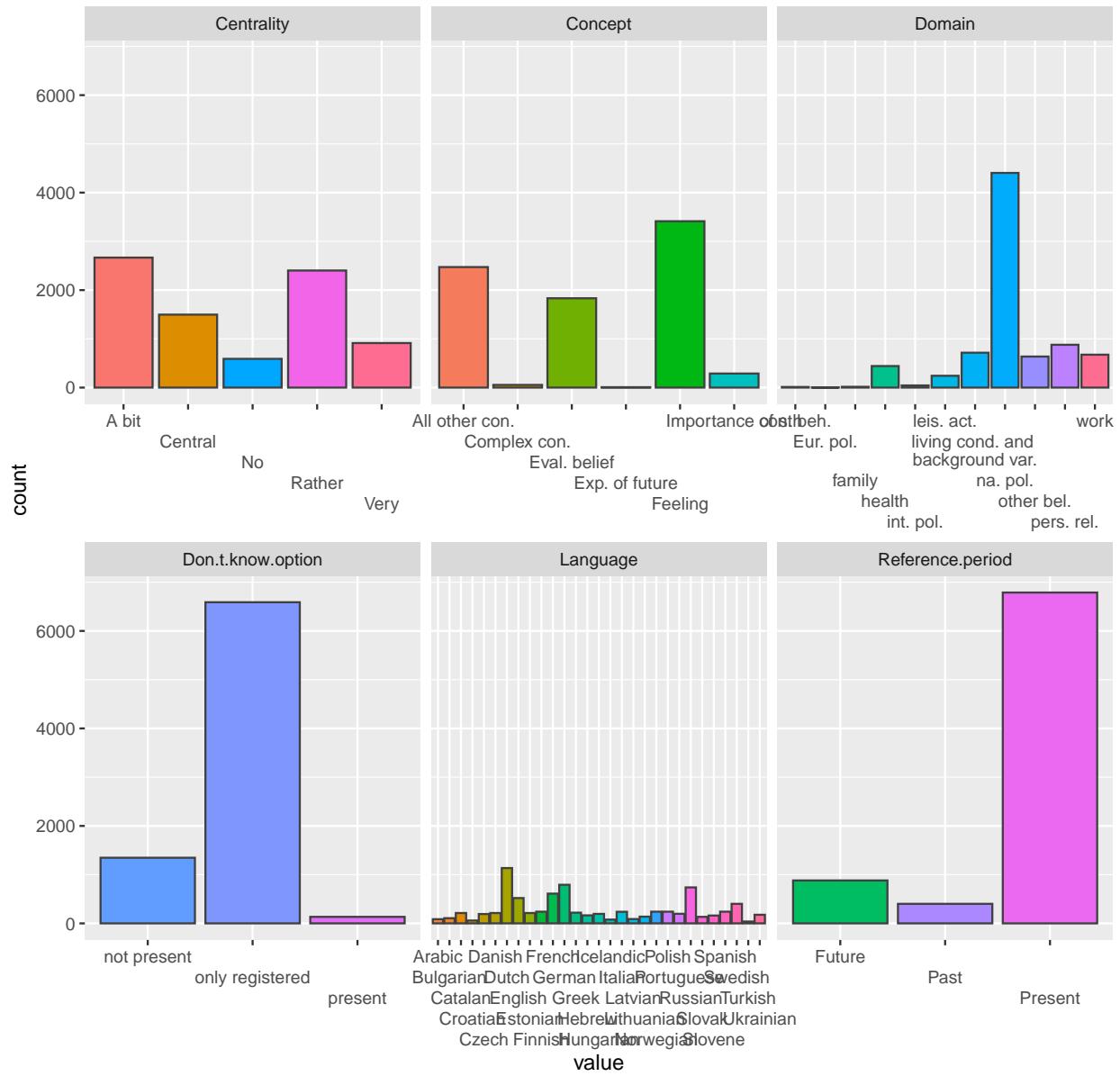


Figure 2: Nicht binäre Einflussgrößen, welche nicht in Filtern vorkommen

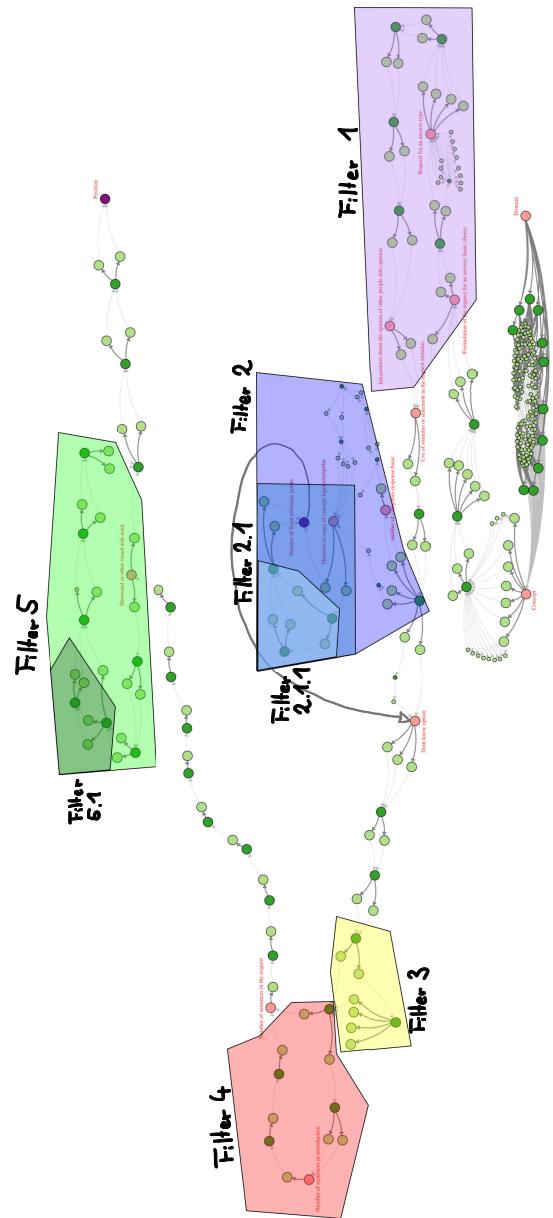


Figure 3: Übersicht über codebook routing mit eingezeichneten Filtern

### 2.3.1 Erster Filter

Der erste Filter ist **Formulation of the request for an answer basic choice**. Nachdem diese Spezifikation mit **indirect request** oder **direct request** “beantwortet” wurde, werden mehrere weitere Spezifikationen zu den **requests** abgefragt.

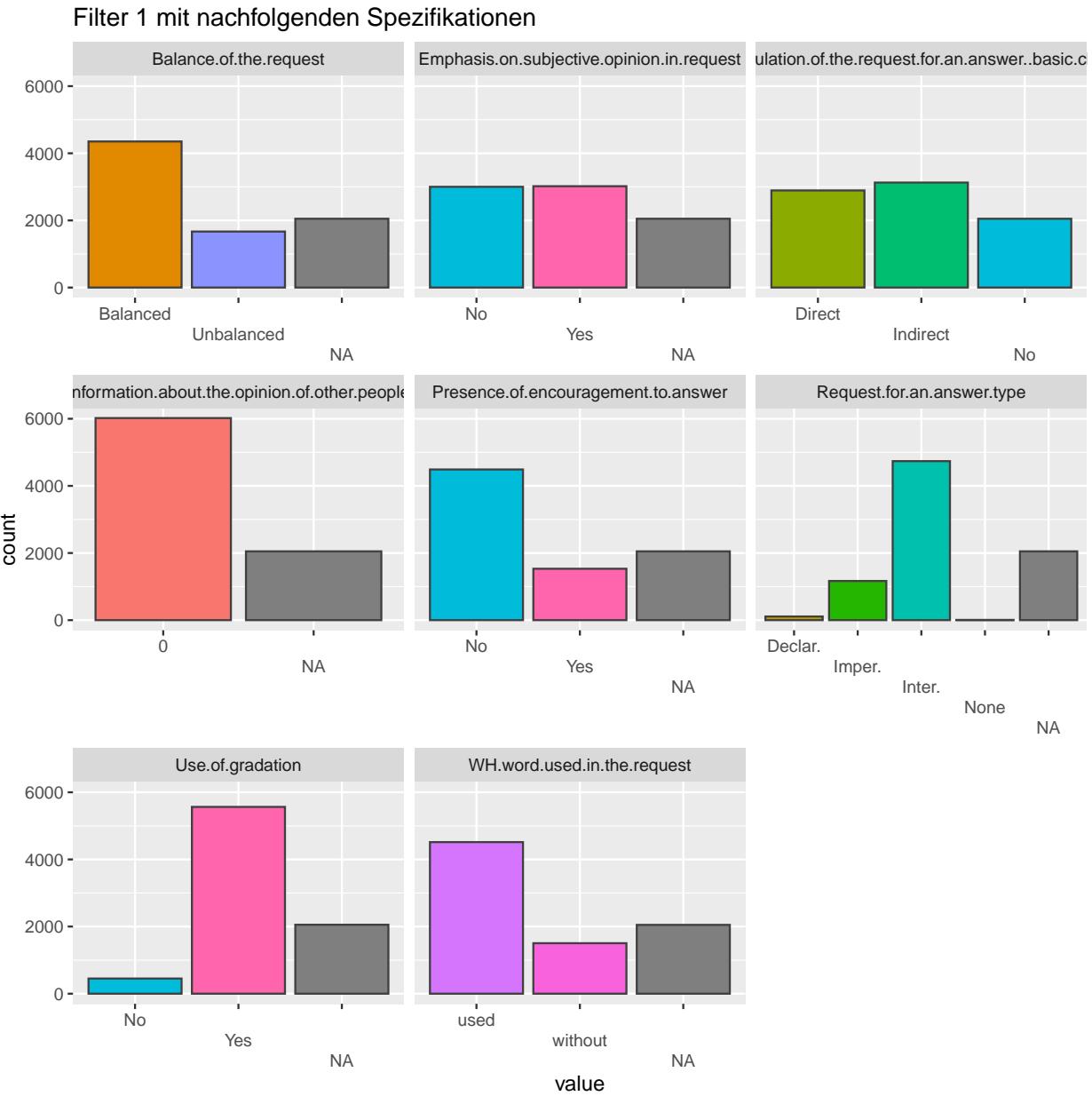


Figure 4: Variablen innerhalb des ersten Filters

Genauere Übersicht über das *Routing*:

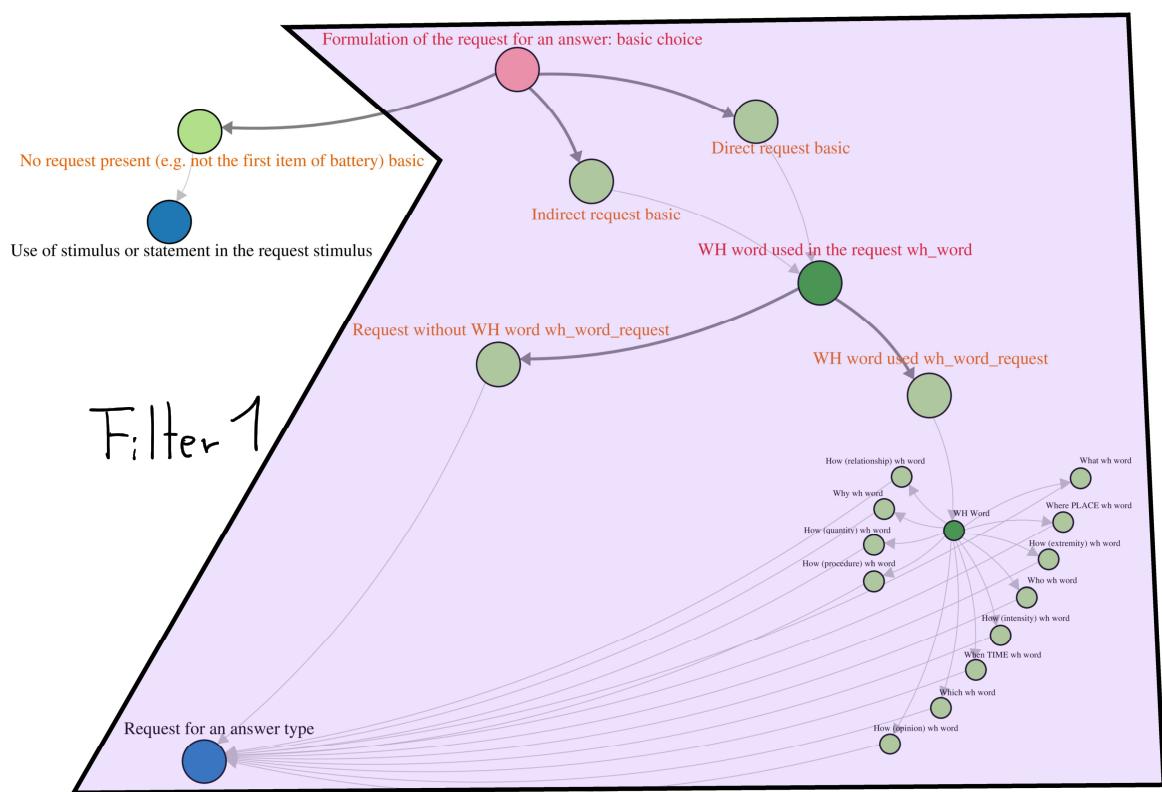


Figure 5: Akkurate Übersicht über den ersten Filter, Teil 1

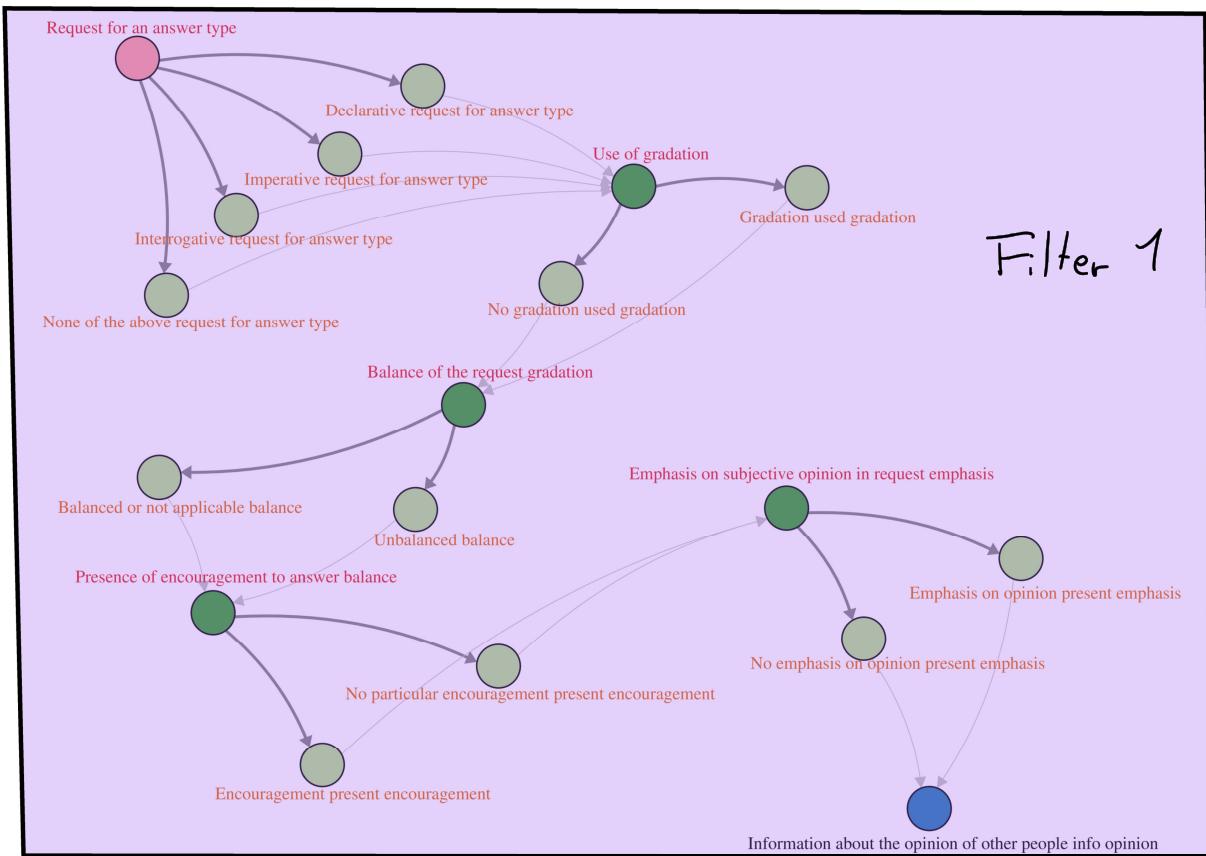


Figure 6: Akkurate Übersicht über den ersten Filter, Teil 2

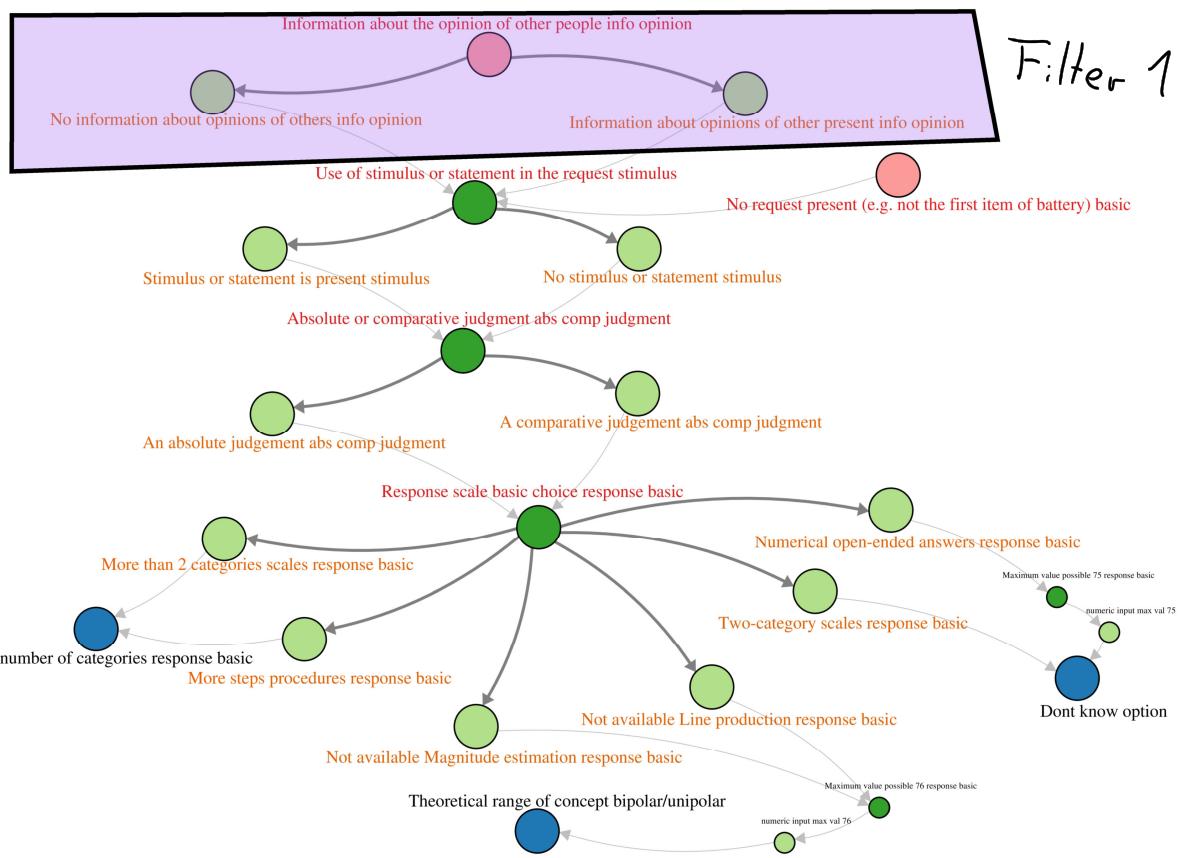


Figure 7: Akkuratere Übersicht über den ersten Filter, Teil 3

### 2.3.2 Zweiter Filter

Der zweite Filter ist **Response scale basic choice**. Nachdem diese Spezifikation mit **More than 2 categories scalec**, **More steps procedures**, **Magnitude estimation** oder **Line production** “beantwortet” wurde, werden mehrere weitere Spezifikationen zu den **response scales** abgefragt.

Diese Filter teilt sich jedoch später noch weiter auf mithilfe vom Filter **Theoretical range of concept bipolar / unipolar** mittels “Antwort” **Theoretically bipolar** und dieser Filter... Filter... spaltet sich nochmals in **Range of the used scale bipolar / unipolar** mittels **Bipolar** auf.

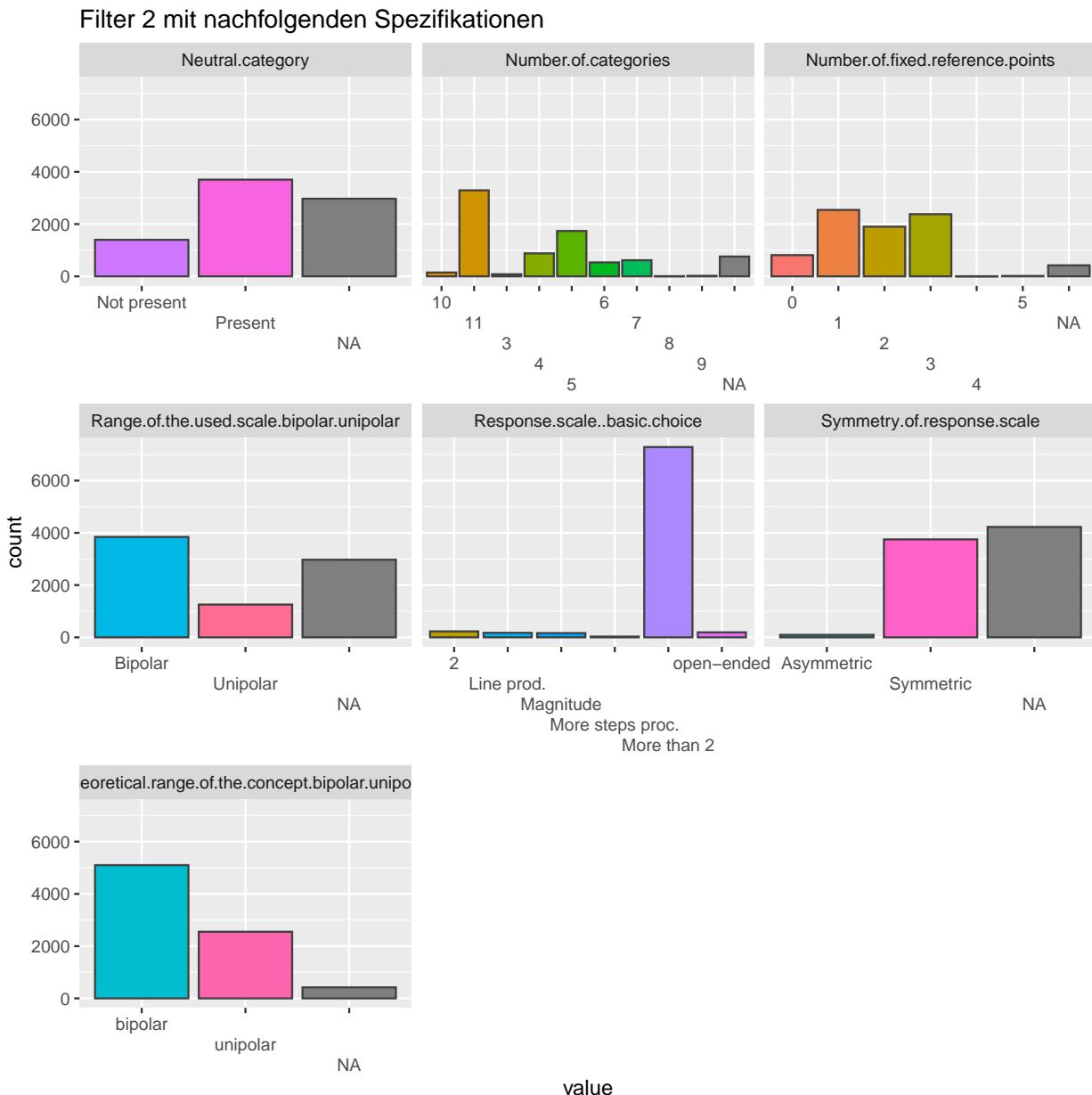


Figure 8: Variablen innerhalb des zweiten Filters

Genauere Übersicht über das *Routing*:

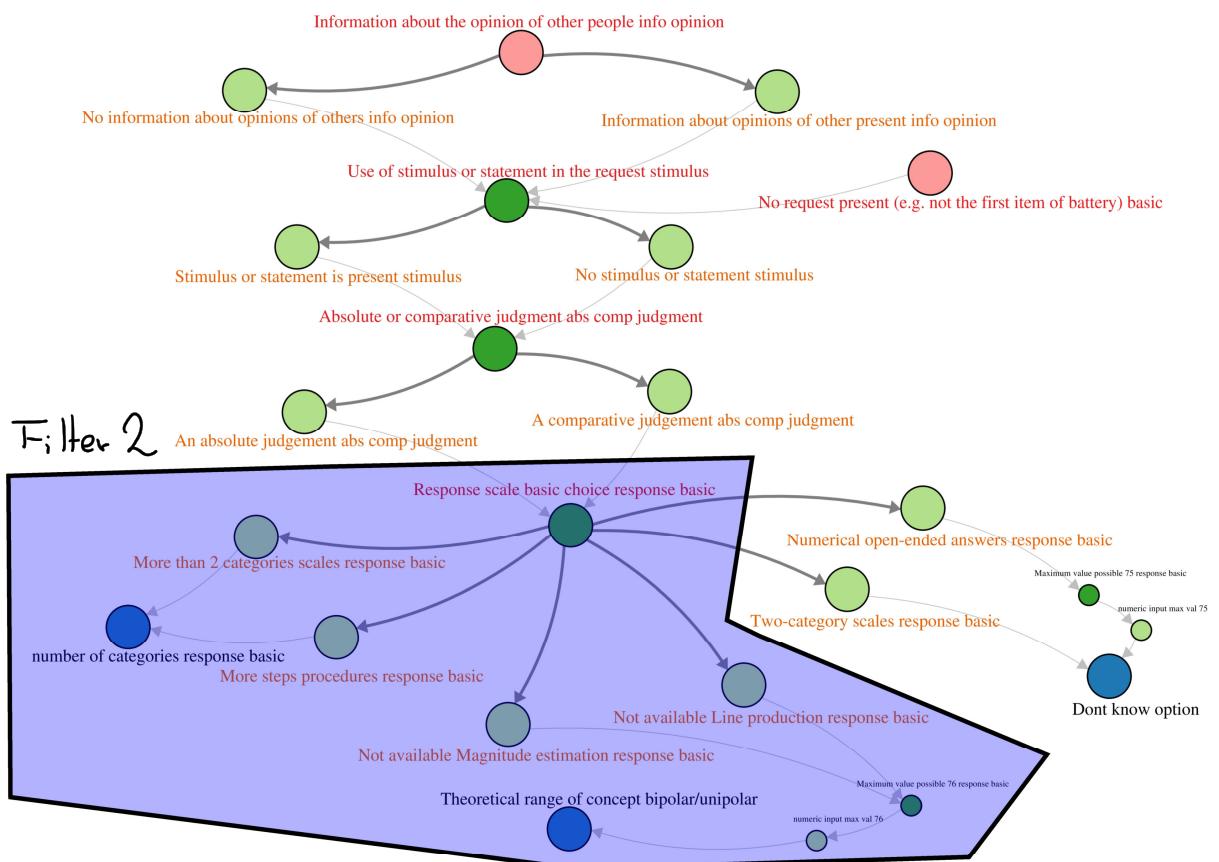


Figure 9: Akkurate Übersicht über den zweiten Filter, Teil 1

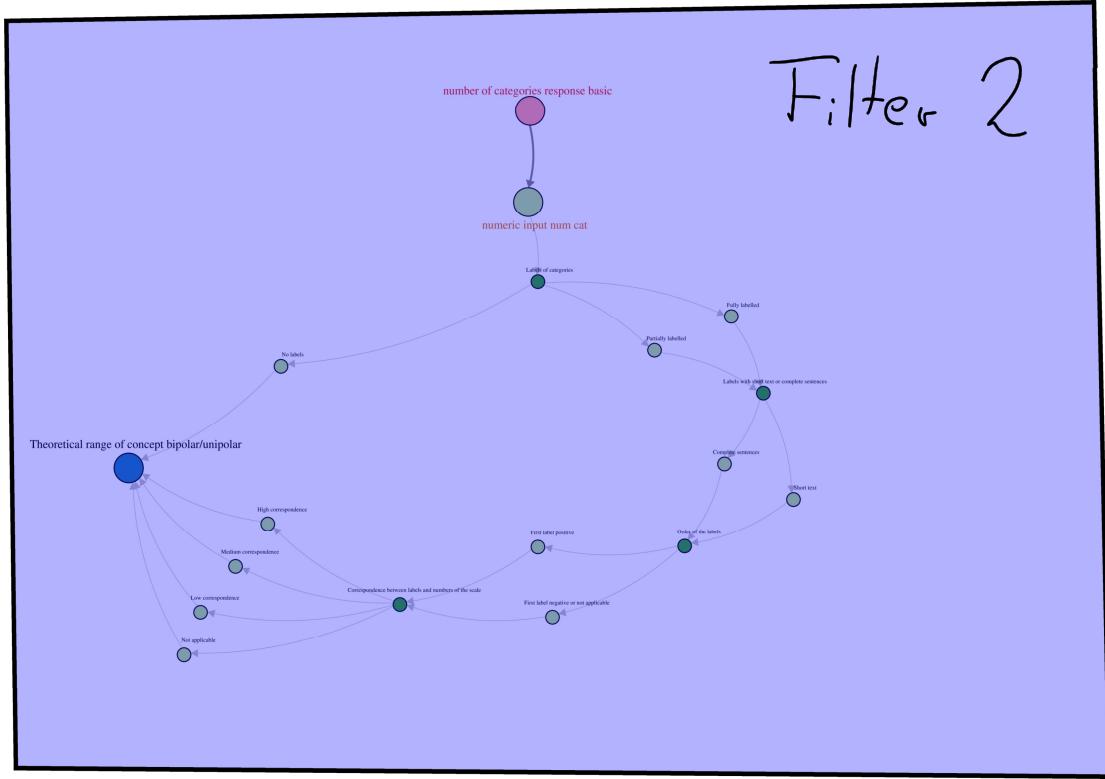


Figure 10: Akkurate Übersicht über den zweiten Filter, Teil 2

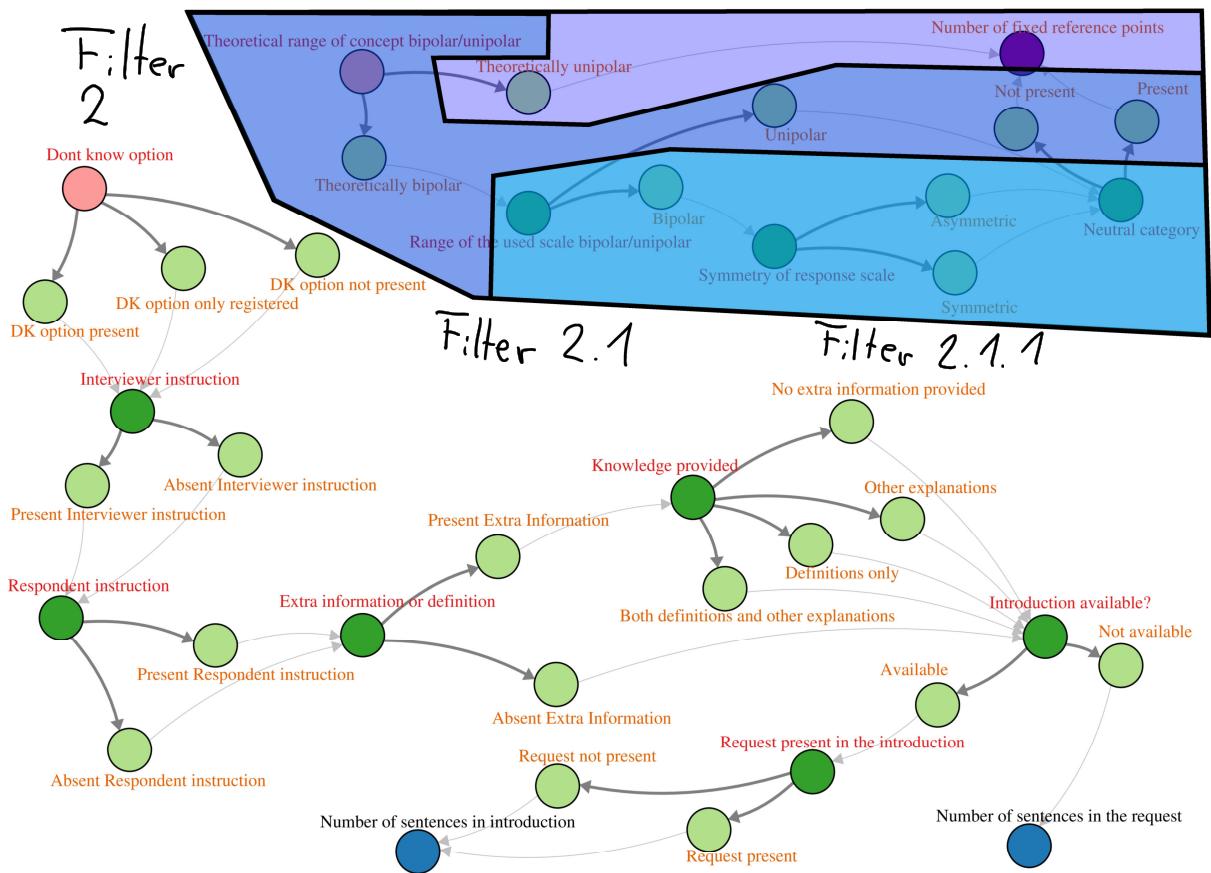


Figure 11: Akkuratere Übersicht über den zweiten Filter, Teil 3

### 2.3.3 Dritter Filter

Der dritte Filter ist **Extra information or definition**. Nachdem diese Spezifikation mit **Present Extra information** „beantwortet“ wurde, liegt in diesem „Filternnpfad“ lediglich **Knowledge provided**.

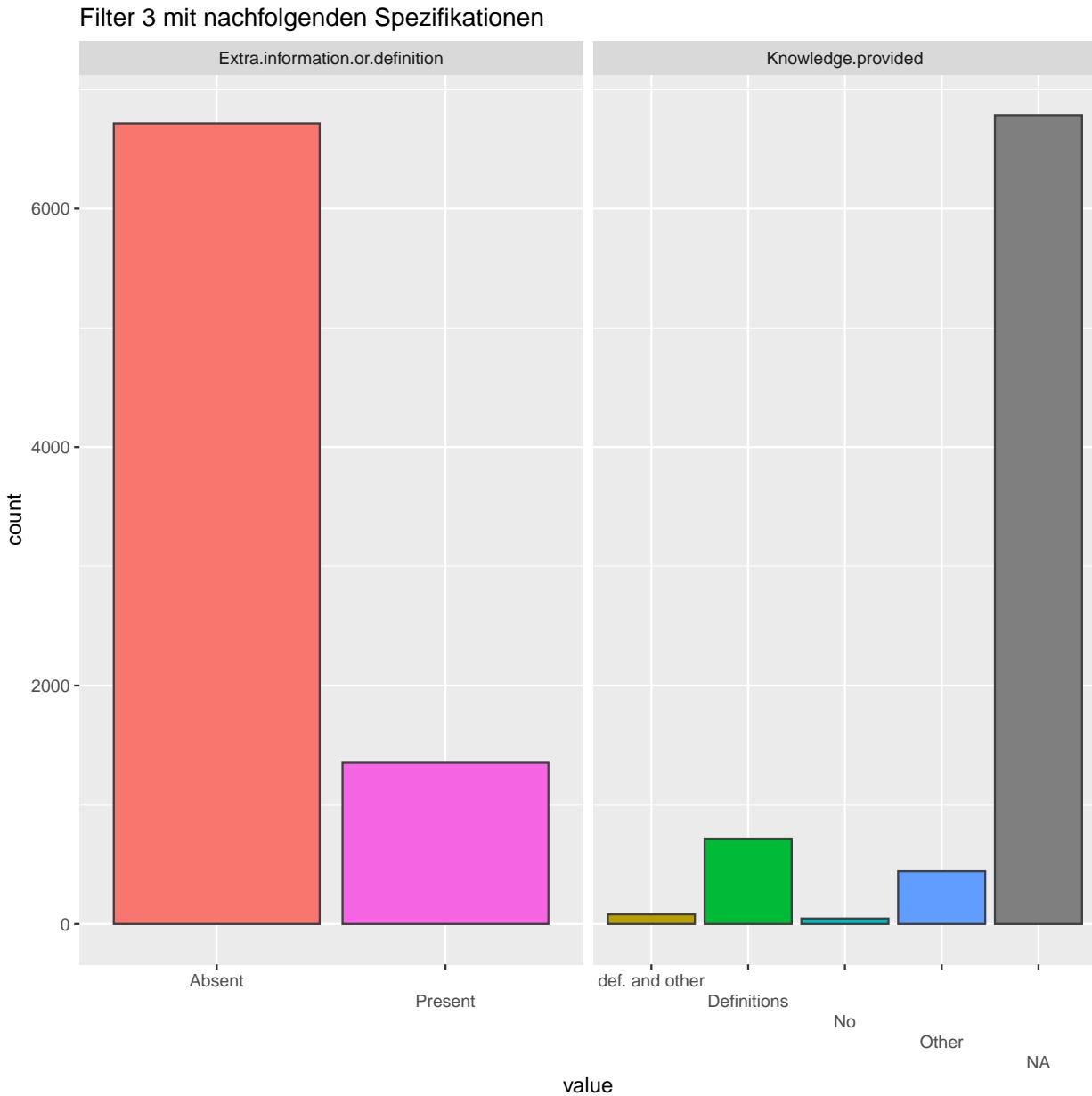


Figure 12: Variablen innerhalb des dritten Filters

Genauere Übersicht über das *Routing*:

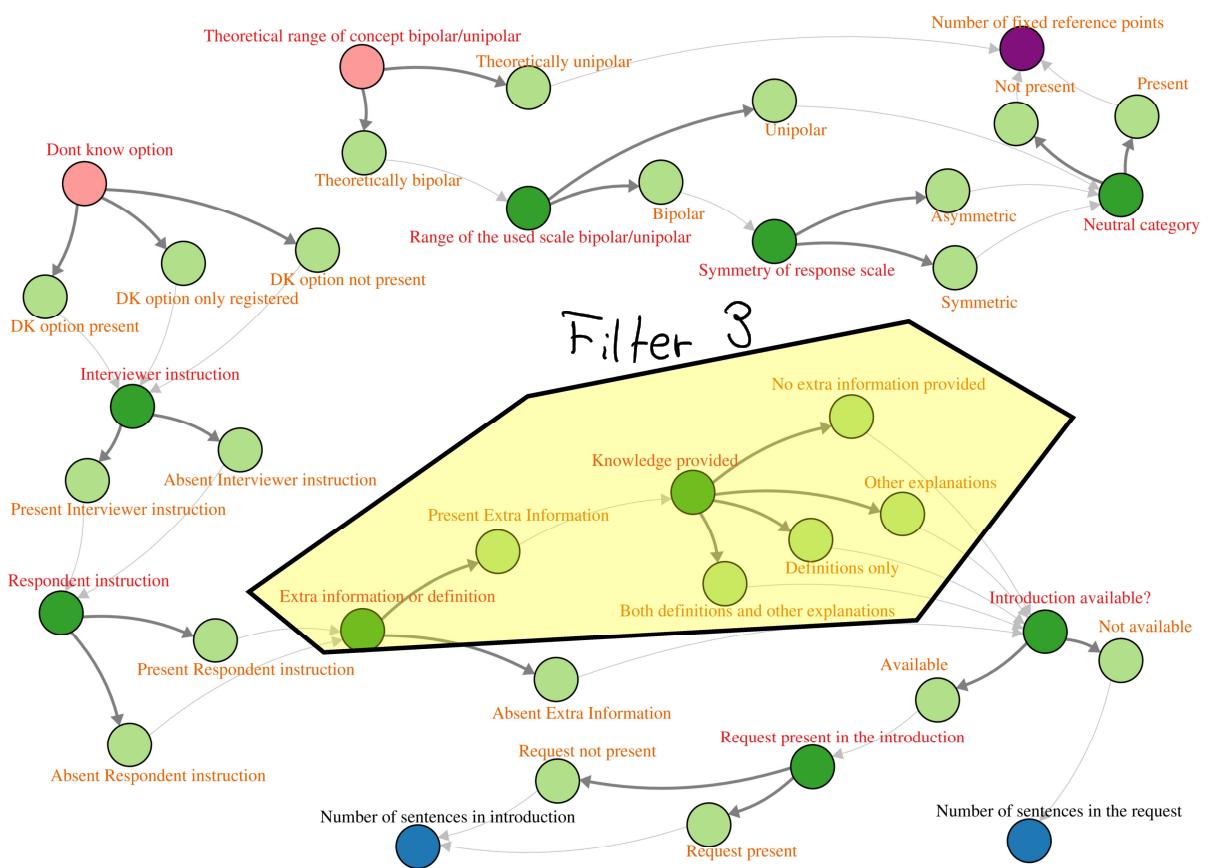


Figure 13: Akkurate Übersicht über den dritten Filter

### 2.3.4 Vierter Filter

Der vierte Filter lautet **Introduction available**. Nachdem diese Spezifikation mit **Available** "beantwortet" wurde, werden mehrere weitere Spezifikationen zur **Introduction** abgefragt.

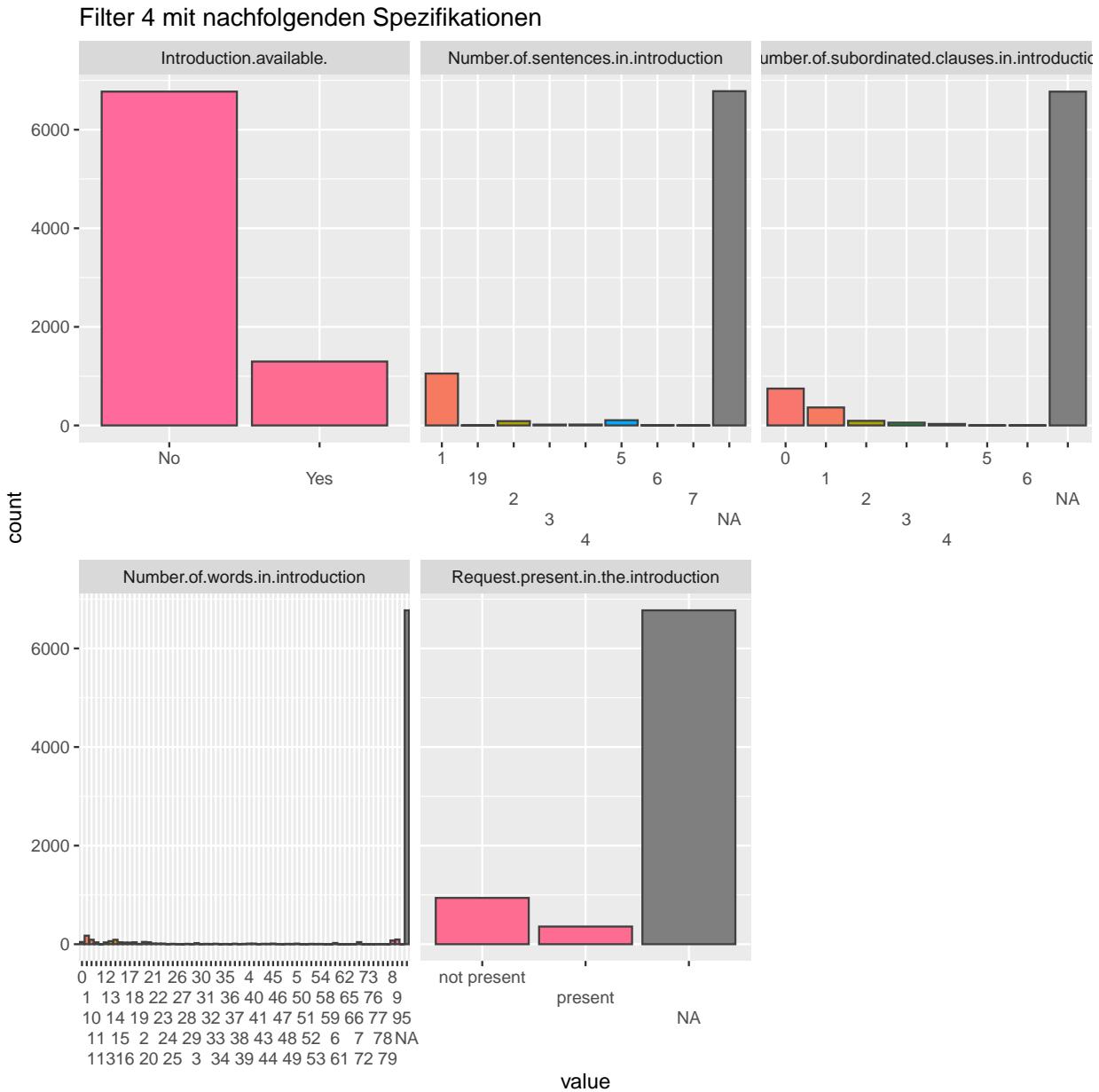


Figure 14: Variablen innerhalb des vierten Filters

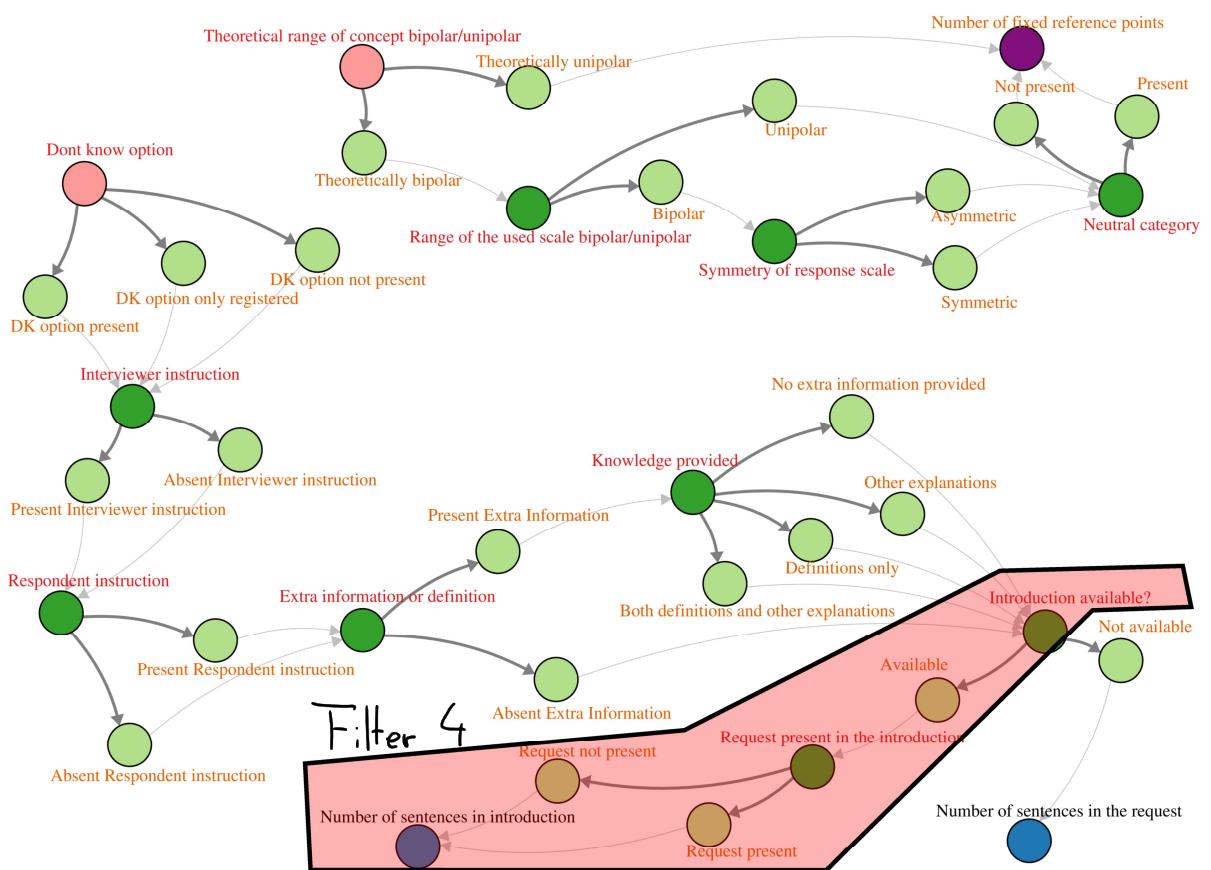


Figure 15: Akkuratere Übersicht über den vierten Filter, Teil 1

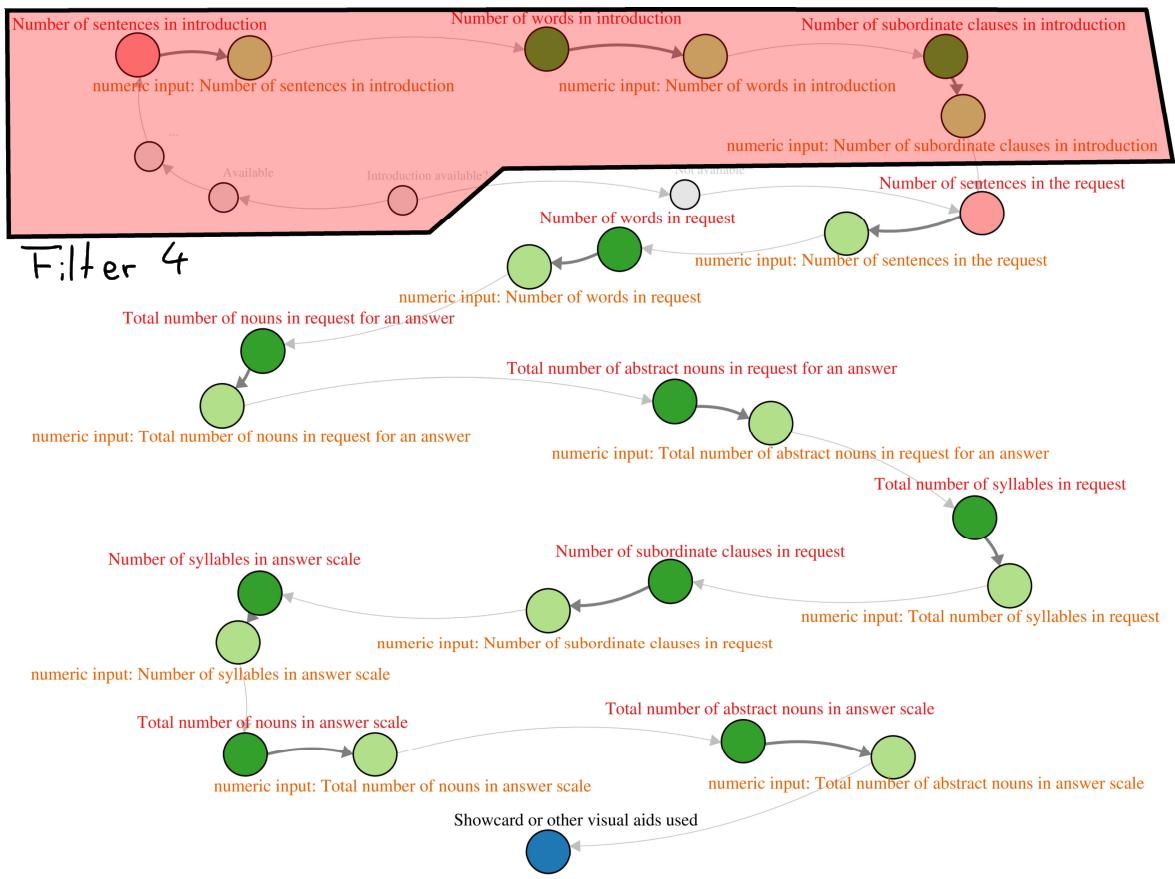


Figure 16: Akkurate Übersicht über den vierten Filter, Teil 2

### 2.3.5 Fünfter Filter

Der fünfte Filter lautet **Showcard or other visual aids**. Nachdem diese Spezifikation mit **Used showcard** "beantwortet" wurde, werden mehrere weitere Spezifikationen zu diesen abgefragt.

Ähnlich der Filter 2 teilt dieser Filter sich mithilfe vom Filter **Numbers or letters before the answer categories** mittels **numbers** oder **letters** auf.

Filter 5 mit nachfolgenden Spezifikationen

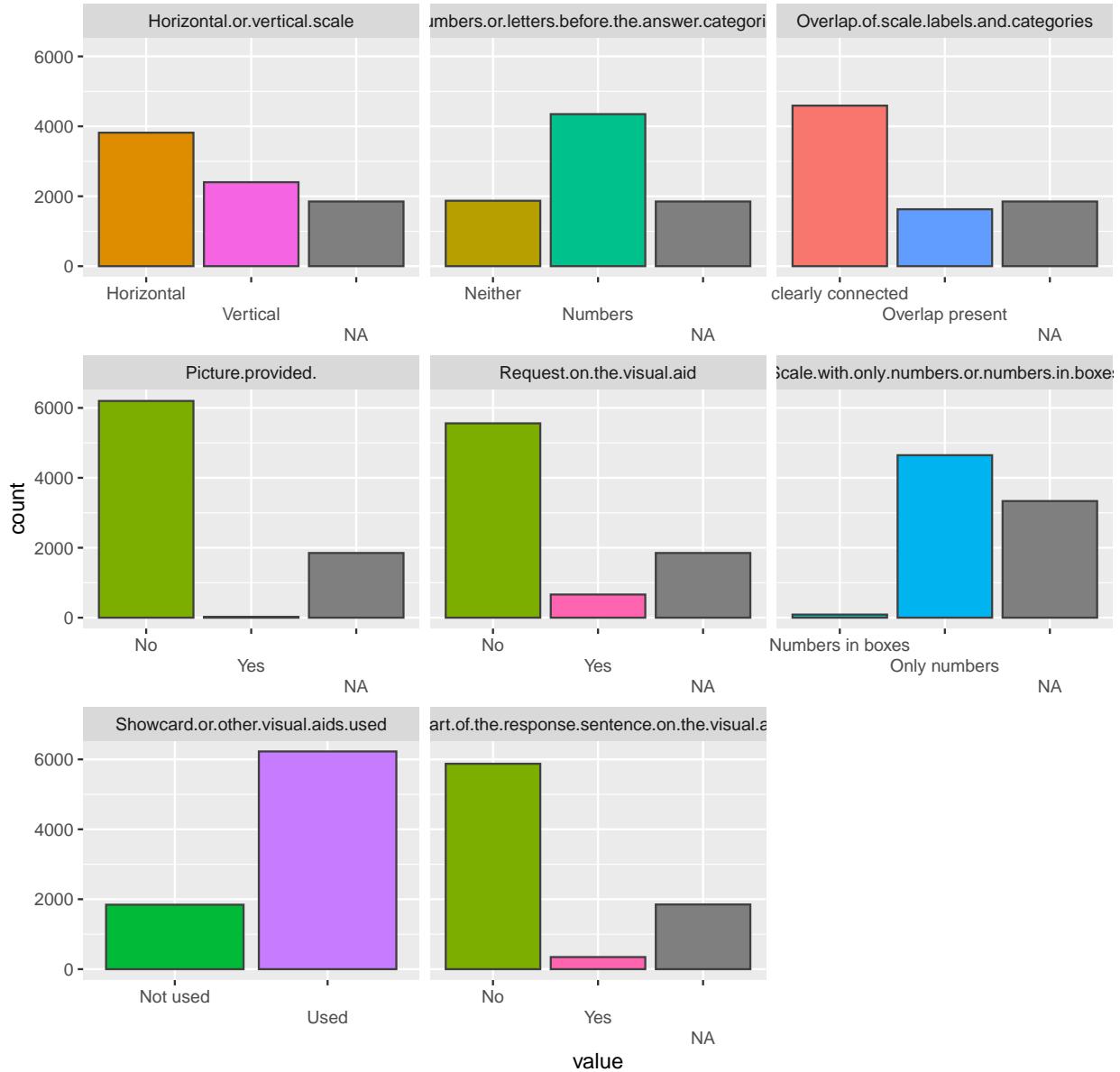


Figure 17: Variablen innerhalb des fünften Filters

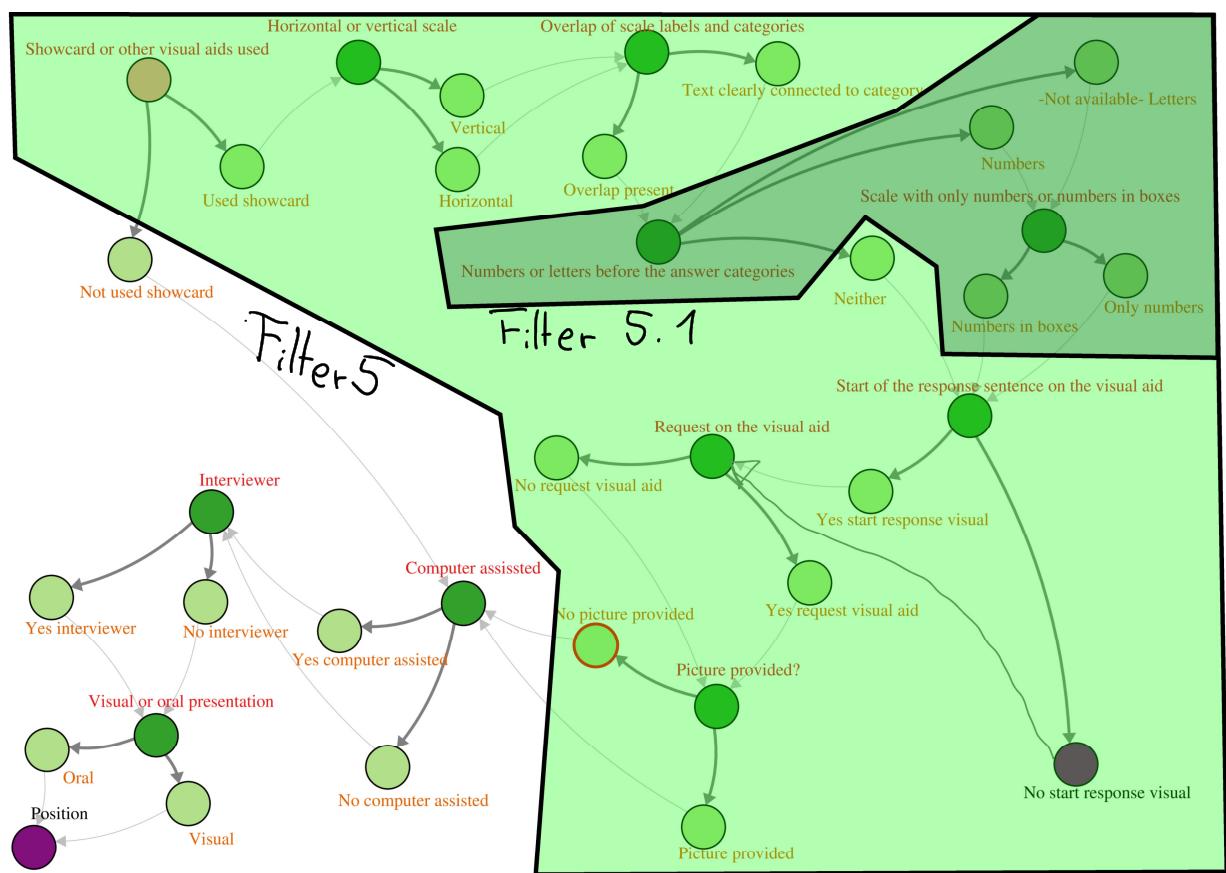
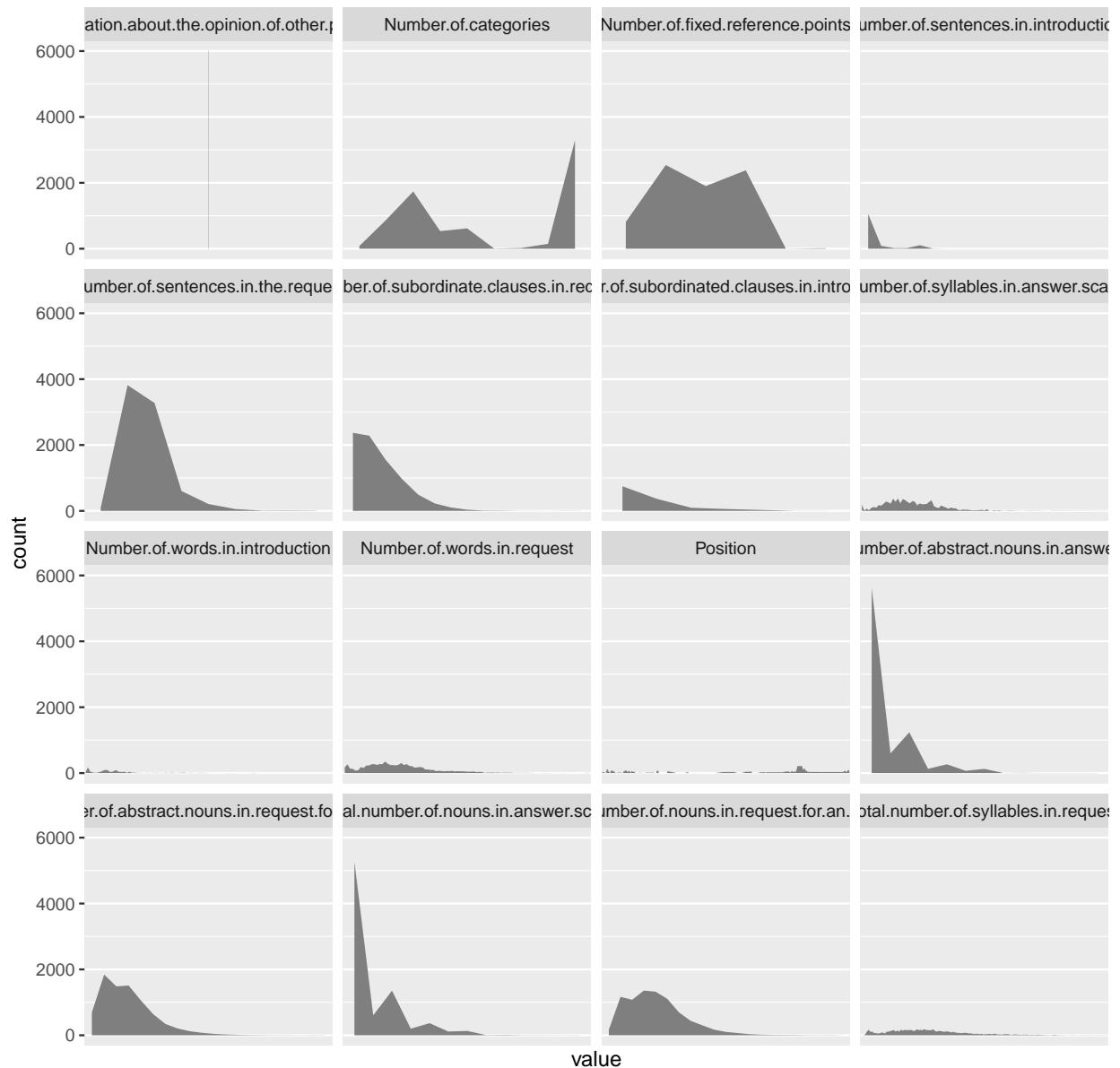


Figure 18: Akkurate Übersicht über den fünften Filter

## 2.4 Metrisch Skalierte Einflussgrößen

Im Gegensatz zu den Nominal -/ Ordinalskalierten Einflussgrößen gibt es wenige **metrischskalierte** Einflüsse (**Anteil** von **0.27**), welche zusätzlich meist in  $\mathbb{N}$  leben.

Übersicht über alle metrischen Einflussgrößen



## 2.5 Outcome: Qualität

Wie oben beschrieben setzt sich die Qualität aus der Reliabilität und der Validität zusammen. Im folgenden Scatterplot kann man erkennen, dass für die Validität und die Reliabilität nur bestimmte (“diskrete”) Werte angenommen werden, was sich bei Werten  $> 0.75$  der Qualität auch bemerkbar macht.

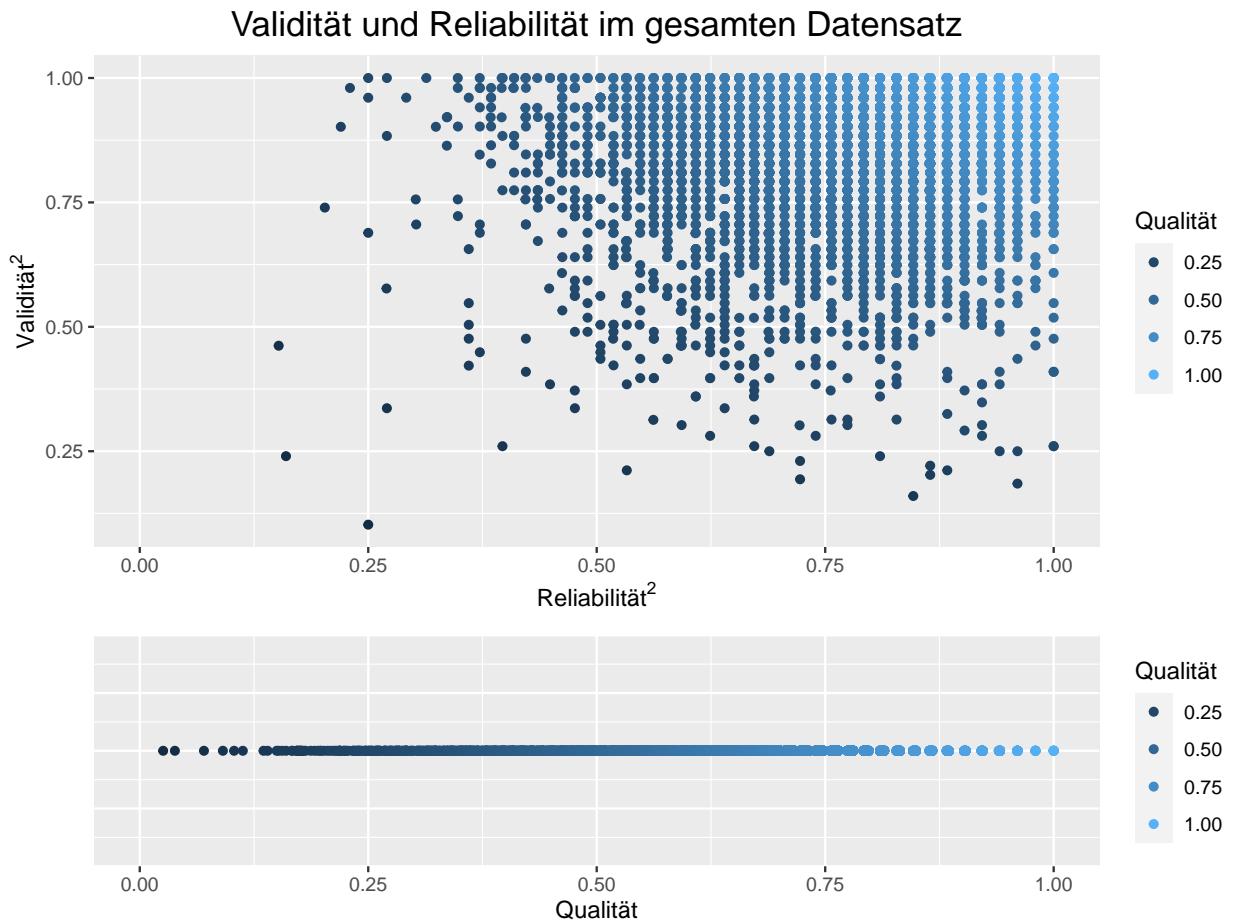


Figure 19: Verteilung der Outcome Variable, sowie den zugehörigen Einflüssen Reliabilität und Validität

### 3 Modelle

Ziel der Modelle soll sein, dass diese interpretierbar sind. Deshalb haben wir uns für Regressionsmodelle entschieden.

Da es viele Beobachtungen gibt (s.o.) kann ein Modell - im Optimalfall - nach der Daumenregel *eine Einflussvariable pro 10 Beobachtungen* insgesamt  $\sim 600$  Variablen beinhalten.

Die hierarchische Struktur der Daten sollte für die Einflüsse **Studien**, **Experimente** und **Sprache** beibehalten werden. Diese verhalten sich folgendermaßen:

#### Studien

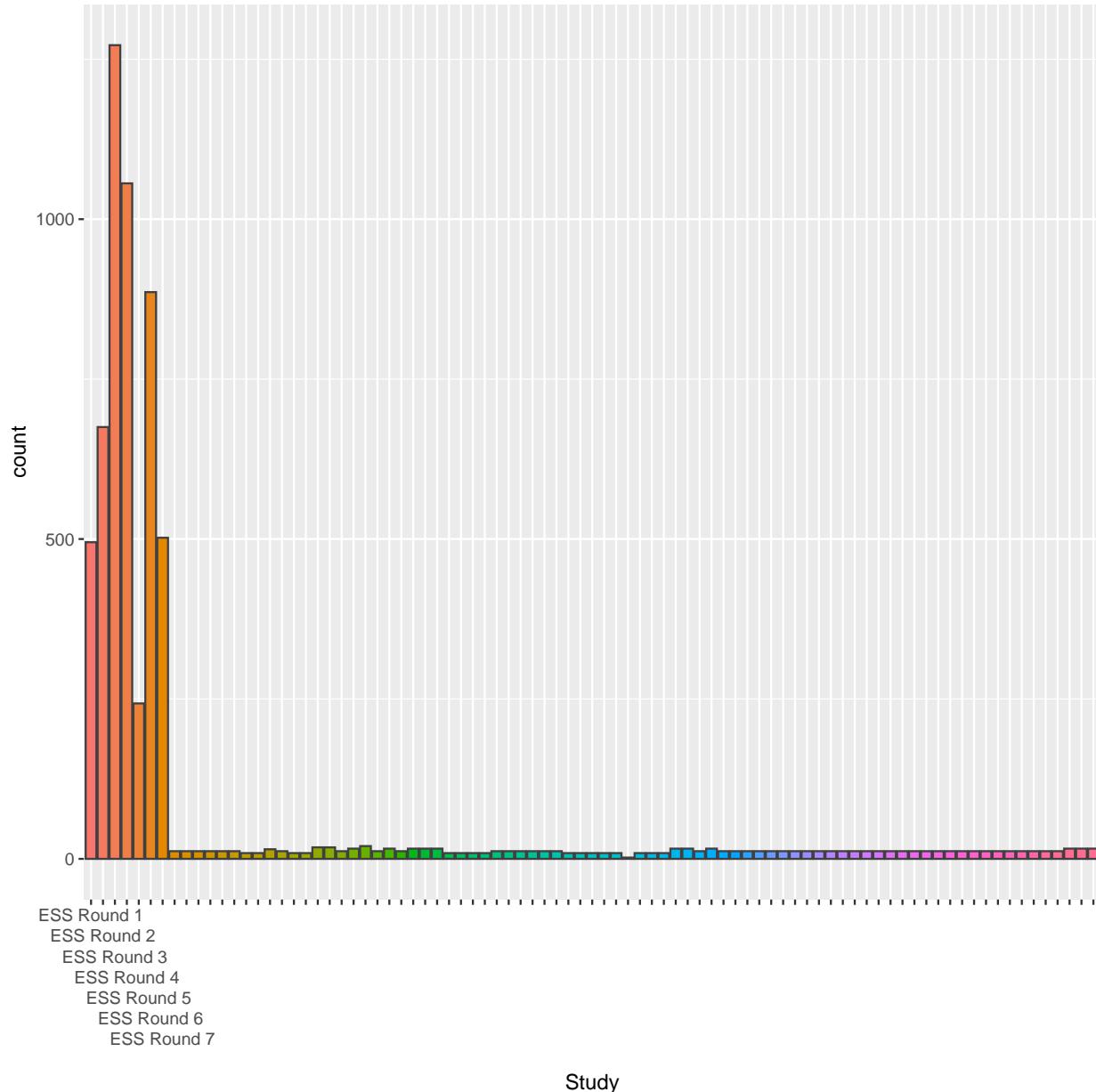


Figure 20: Häufigkeitsverteilung der Studien

Es ist erkennbar, dass einige wenige Studien (8) (ausschließlich ESS) auf 5129 Beobachtungen kommen. Somit

wäre es eventuell sinnvoll alle restlichen Studien als “restliche” zusammenzufassen mit 945 Beobachtungen. Daraus resultieren 9 Studien.

## Sprachen

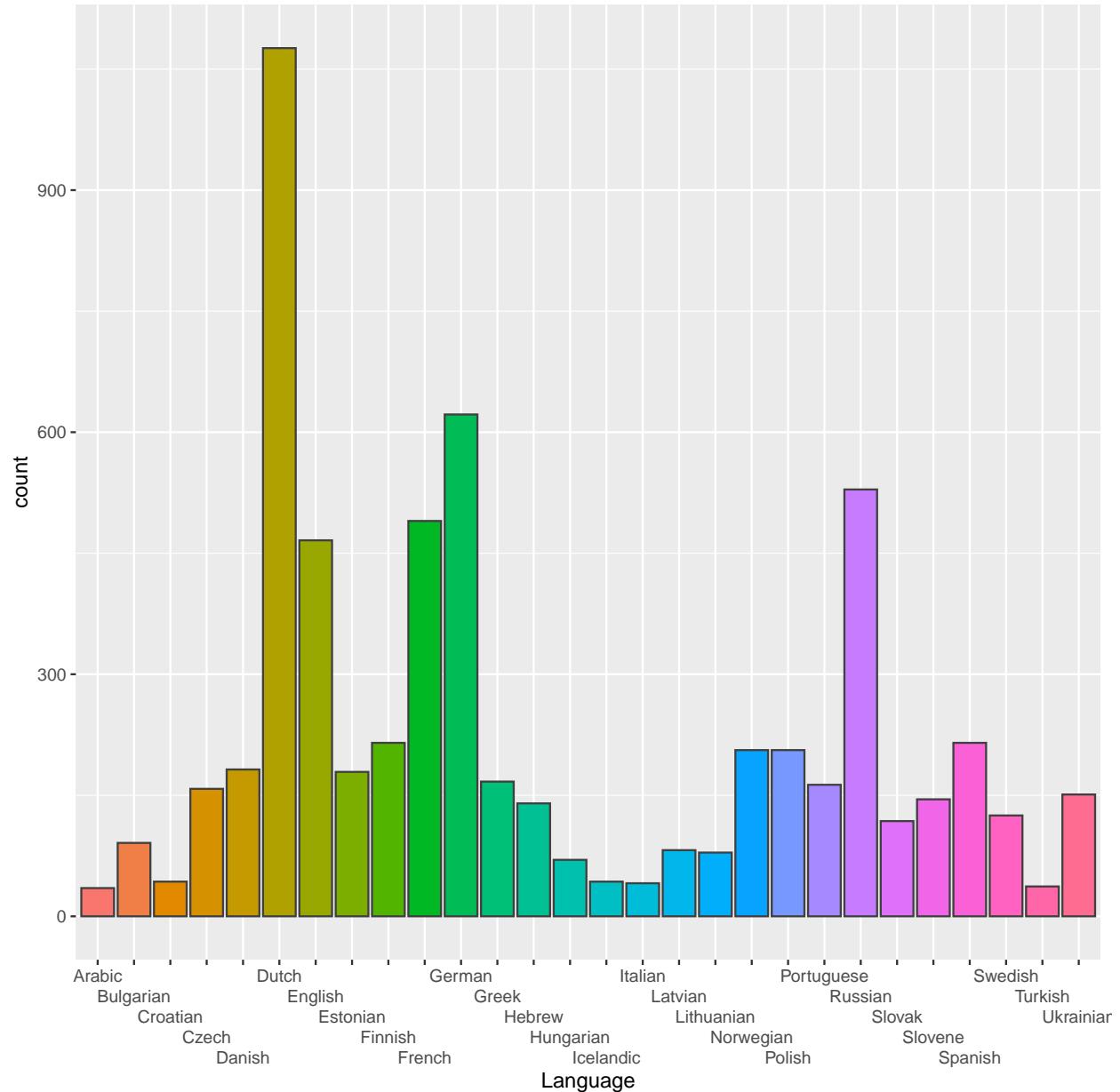


Figure 21: Häufigkeitsverteilung der Sprachen

Im Gegensatz zu den Studien scheinen die Sprachen gleichverteilter zu sein. Jedoch gibt es hierbei 5 Sprachen, welche über 300 mal auftreten. Der Anteil dieser Sprachen macht 0.52 der Beobachtungen aus.

### 3.1 Genestete Variablen

Im Modell soll später berücksichtigt werden, dass eine hierarchische Struktur der Daten existiert. Diese ist wie folgt aufgebaut: Studien in Experimenten in Ländern. Diese sollten heterogen verteilt sein, d.h. es sollte beispielsweise für **ESS 1** nicht nur ein Land repräsentiert sein.

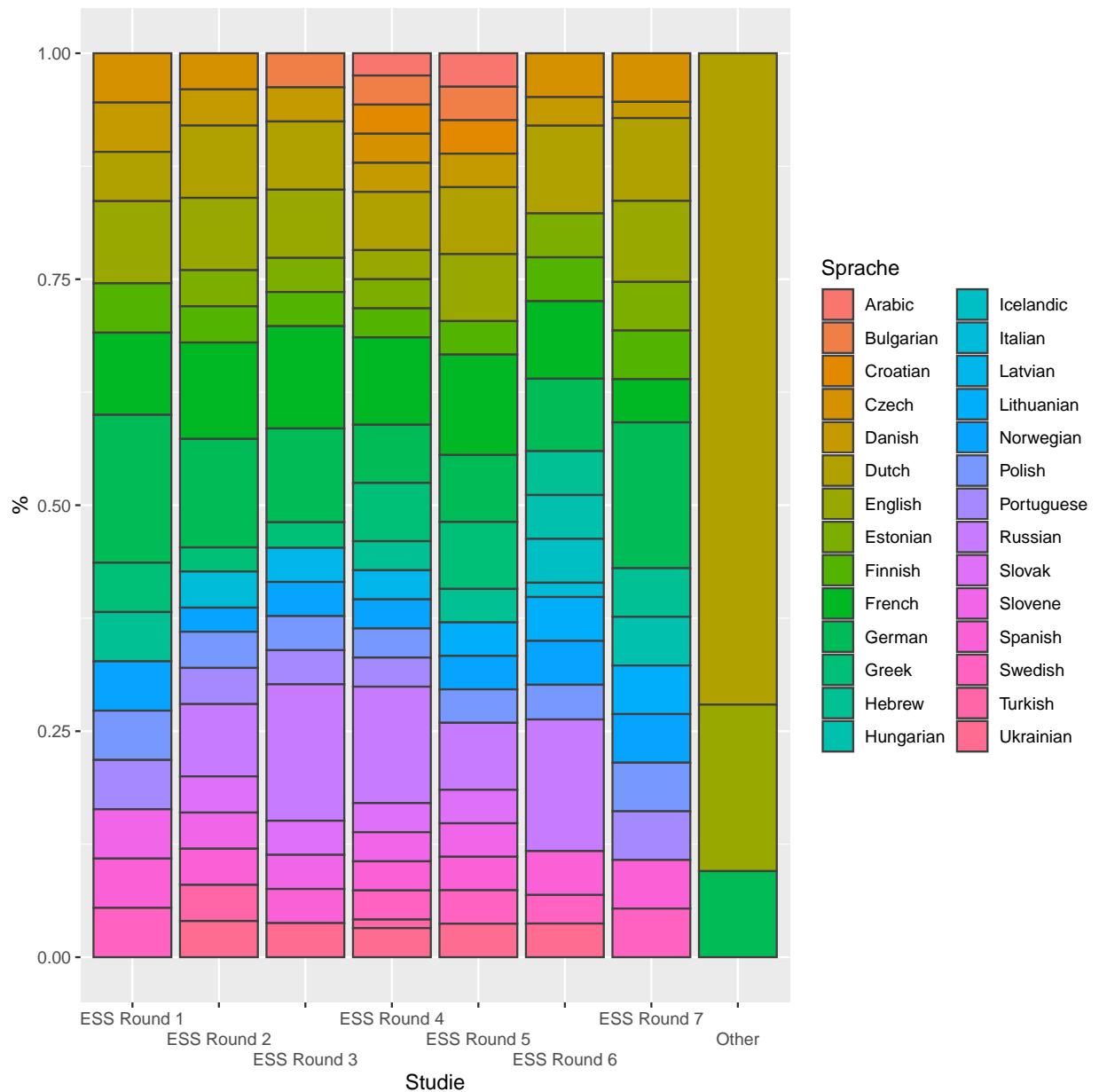


Figure 22: Verteilung der Länder in Studien

In der oberen Grafik erscheinen die Länder in den **ESS Studien** gleichverteilt zu sein. Eine Ausnahme sind hierbei die **Other** Studien, welche vor allem in der Niederlande gemacht wurden.

Um die Gleichverteilung ein wenig näher zu betrachten werden im folgenden die Lorenzkurven der einzelnen Studien visualisiert.

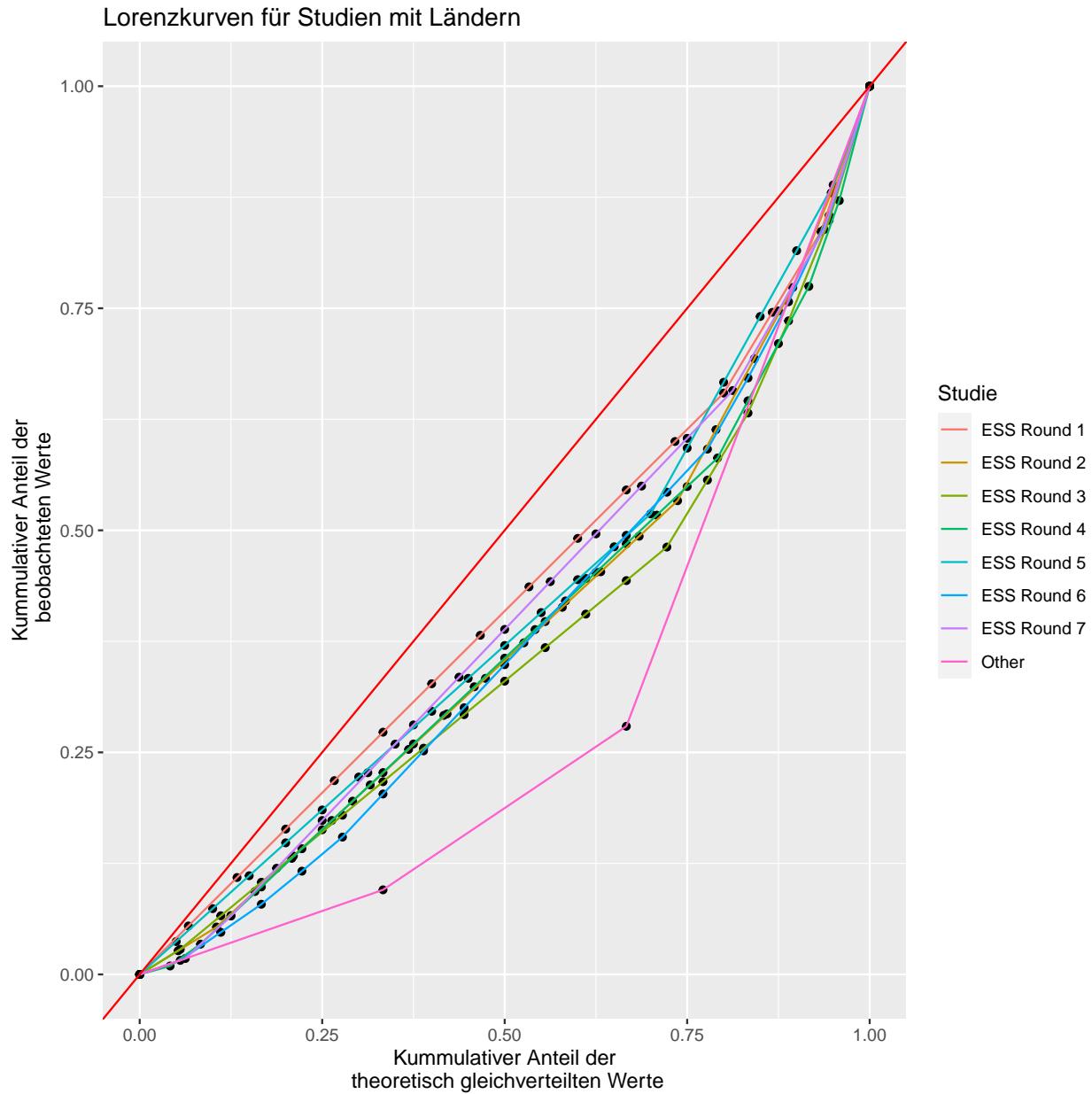


Figure 23: Lorenzkurven bezüglich der Häufigkeit von Ländern, separat berechnet für jede Studie

Wie in der vorherigen Grafik scheinen die *ESS Studien* gleichverteilt zu sein, jedoch sind - im Gegensatz zu den *ESS Studien* - die *Other* Studien ungleich verteilt.

Dies spiegelt sich auch in den einzelnen Gini-Koeffizienten wieder:

Table 3: Gini-Koeffizient bezüglich der Häufigkeit von Ländern berechnet für jede Studie

Gini	Study
0.155	ESS Round 1
0.232	ESS Round 2
0.268	ESS Round 3
0.246	ESS Round 4
0.191	ESS Round 5
0.248	ESS Round 6
0.195	ESS Round 7
0.417	Other

## 3.2 Modelle

Nach bisherigem Stand haben wir die Möglichkeit zwei verschiedene Modelle zu berechnen. Beide male sollte ein LMM oder ein GLMM (je nach Verteilung) berechnet werden, wobei Cluster durch die Sprache und / oder durch Sprache mit der jeweiligen Studie entstehen. Jedoch besteht noch immer das Problem, dass Modelle nicht berechnet werden können aufgrund der Filter. Hier zwei Methoden um damit umzugehen:

### 3.2.1 Modell 1: Mit Einflussvariablen umkodieren

Eine Möglichkeit zum Umgang mit den Filtern bietet das Paper *Modeling with Structurally Missing Data by OLS and Shapley Value Regressions* von *Stan Lipovetsky und Ewa Nowakowska*.

“[...] we suggest splitting each  $x_j$  with missing values into a system of binary variables, or Gifi system. If  $x_j$  is measured by a several-point Likert scale, each binary variable of a Gifi system identifies each of the levels of the scale by using the value of 1, and all the other levels including NA equal zero. Only the binary variables of the numerical levels are needed, and the missing values serve as a reference.”

Somit könnte man alle Träger der Filter als eine eigene binäre Variable umkodieren.

### 3.2.2 Model 2: Eigene Strata für die Filter

Die zweite Möglichkeit, die in der Einleitung im Buch *Statistical Analysis with Missing Data (second edition)* von *Roderick Little und Donald Rubin* (S.8) steht:

“We make the following key assumption throughout the book:

#### **Assumption 1.1: missingness indicators hide true values that are meaningful for analysis**

Assumption 1.1 may seem innocuous, but it has important implications for the analysis. When the assumption applies, it makes sense to consider analyses that effectively predict, or “impute” (that is, fill in) the unobserved values. If, on the other hand, Assumption 1.1 does not apply, then imputing the unobserved values makes little sense, and an analysis that creates strata of the population defined by the missingness indicator is more appropriate.”

Da die Filter mit den NAs keinen wahren Wert “verstecken”, sondern dies so strukturell aufgebaut wurde, sollten somit die Strata betrachtet werden. Hierbei müsste für jede Möglichkeit der Filter ein eigenes Modell berechnet werden. Problematisch wäre die geringe Anzahl an Beobachtungen und noch weiter kann ich nicht sagen, ob überhaupt bei so wenigen Beobachtungen die Unterteilung per LMM / GLMM (aufgrund der hierarchischen Struktur) sinnvoll wäre. Hier fehlen noch Zahlen!

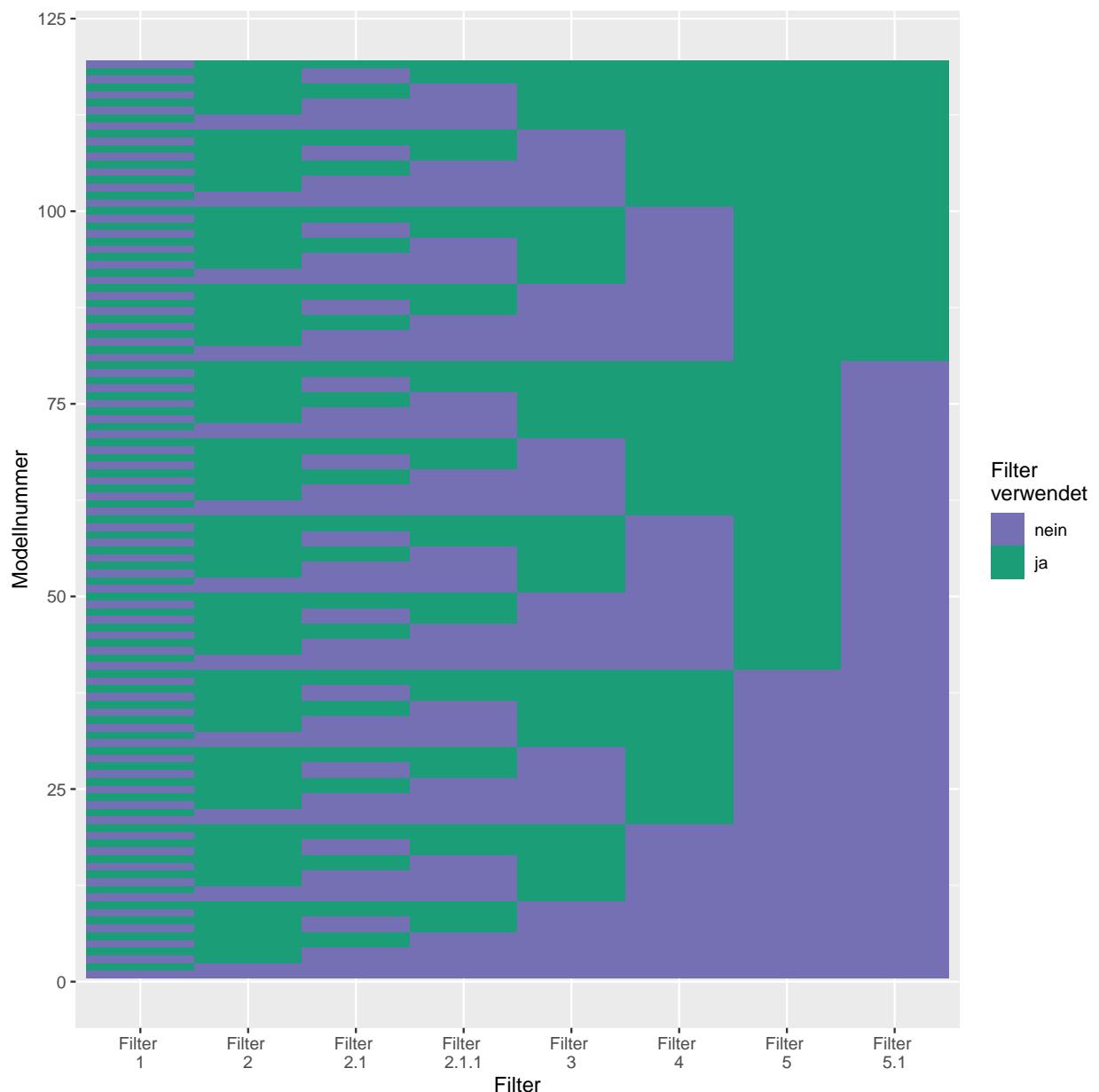


Figure 24: Alle Modelle, falls für alle Strata / Filter ein Modell berechnet wird