



Hochschule für Angewandte Wissenschaften Hamburg  
*Hamburg University of Applied Sciences*

DEPARTMENT INFORMATION

## ***Bachelorarbeit***

Relevanzanalyse niedrig geranker Google-Suchergebnisse  
auf informationsorientierte Suchanfragen mit dem Relevan-  
ce Assessment Tool

vorgelegt von

***Lea Podgajnik***

Studiengang Medien und Information

Hochschule für Angewandte Wissenschaften Hamburg  
Fakultät Design, Medien & Information  
Department Information  
Finkenau 35  
22081 Hamburg

Verfasser: Lea Podgajnik  
Abgabedatum: 10.10.13

Erster Prüfer: Prof. Dr. Dirk Lewandowski  
Zweite Prüferin: Prof. Dr. Ulrike Spree

Hamburg, Oktober 2013

**Abstract:**

Die Websuche stellt eine der meistgenutzten Internetanwendungen dar und nimmt in ihrer Bedeutung bei der Informationsgewinnung sowohl im professionellen als auch alltäglichen Bereich, stetig zu. Umso wichtiger ist die Frage nach dem Ranking der Treffer und der Qualität der Suchergebnisse, die Suchmaschinen wieder Marktführer Google liefern.

Gegenstand dieser Arbeit ist die Überprüfung der Retrievaleffektivität, also der Fähigkeit der Suchmaschine auf Suchanfragen relevante Ergebnisse zu liefern und der Qualität des Rankings von Google. Als Methode wird ein Online-Retrievaltest durchgeführt, indem informationsorientierte reale Suchanfragen an eine Suchmaschine gestellt, die zurückgegebenen Ergebnisse der Positionen 1 bis 100 gemischt und einer Gruppe von Juroren in einem Onlinetest zur Relevanzbewertung vorgelegt werden. Dabei wird das Relevance Assessment Tool als Onlinewerkzeug für die Umsetzung und Durchführung von Retrievaltests verwendet.

Nach einer statistischen Auswertung der erhobenen Daten sollen Aussagen über die Retrievaleffektivität der Suchmaschine Google bis auf niedrig gerankte Trefferpositionen und die Beurteilung des Rankingverfahrens möglich sein.

**Schlagwörter:**

Information-Retrieval-Evaluierung, Informationsorientierte Suchanfragen, Internetsuche, Nutzerverhalten, Online-Test , RAT, Relevanz , Retrievaleffektivität, Retrievaltest, Suchmaschine, Suchmaschinenqualität, World Wide Web

# Inhaltsverzeichnis

|         |  |    |
|---------|--|----|
| 1       | Einleitung.....  | 1  |
| 1.1     | Zweck und Abgrenzung der Arbeit.....                                   | 4  |
| 1.2     | Lösungsansatz und Methodik der Arbeit.....                             | 7  |
| 1.3     | Struktur der Arbeit .....  | 9  |
| 2       | Selektions- & Navigationsverhalten der Nutzer auf SERP .....           | 10 |
| 2.1     | Analysemethoden des Selektions- & Navigationsverhalten auf SERP.....   | 10 |
| 2.1.1   | Überwachte Erhebungsmethoden .....                                     | 11 |
| 2.1.1.1 | Fragebögen.....  | 11 |
| 2.1.1.2 | Laborexperimente .....   | 12 |
| 2.1.1.3 | Blickaufzeichnung / Eye Tracking .....                                 | 14 |
| 2.1.1.4 | Online Panels .....  | 15 |
| 2.1.2   | Unüberwachte Erhebungsmethoden.....                                    | 16 |
| 2.1.2.1 | Logfileanalyse.....  | 16 |
| 2.1.2.2 | Livetickeranalyse.....   | 17 |
| 2.1.3   | Methodenüberblick- & Zusammenfassung.....                              | 18 |
| 2.2     | Stand der Forschung zum Selektions- & Navigationsverhalten auf SERP    | 20 |
| 2.2.1   | Studien zu unveränderten SERP .....                                    | 20 |
| 2.2.1.1 | Die Studie von Granka et al. ....                                      | 20 |
| 2.2.1.2 | Die Enquiro-Studien von Hotchkiss et al.....                           | 21 |
| 2.2.1.3 | Die Studie von van Gisbergen et al. ....                               | 25 |
| 2.2.1.4 | Die Studie von González-Caro & Marcos .....                            | 26 |
| 2.2.2   | Studien zu manipulierten SERP .....                                    | 28 |
| 2.2.2.1 | Die Studie von Pan et al. ....   | 28 |
| 2.2.2.2 | Die Studie von Cutrell & Guan .....                                    | 31 |
| 2.3     | Erkenntnisse über das Selektions- & Navigationsverhalten auf SERP..... | 33 |
| 3       | Retrievaltests zur Suchmaschinenevaluierung .....                      | 35 |
| 3.1     | Der Begriff der Relevanz .....   | 35 |
| 3.2     | Methoden und Standards für Retrievaltests .....                        | 36 |
| 3.2.1   | Das Cranfield Modell .....   | 37 |

|   |    |
|---|----|
| 3.2.2 Testaufbau nach TREC .....  | 37 |
| 3.2.3 Konzeption eines Retrievaltests nach Tague-Sutcliffe .....              | 38 |
| 3.2.4 Kriterien für Retrievaltests bei Suchmaschinen nach Hawking et al. .... | 39 |
| 3.2.5 Framework für Retrievaltests nach Lewandowski.....                      | 40 |
| 3.2.6 Maße und Kennzahlen für die Datenauswertung in Retrieval Tests ....     | 43 |
| 3.2.6.1 Die Maße Recall und Precision als Standard in Retrievaltests.....     | 44 |
| 3.2.6.2 Weitere Kennzahlen .....  | 46 |
| 3.3 Stand der Forschung über Retrievaltests .....                             | 47 |
| 3.3.1 Die Retrievaltests von Griesbaum .....                                  | 47 |
| 3.3.2 Der Retrievaltest von Véronis.....                                      | 50 |
| 3.3.3 Der Retrievaltest von Lewandowski .....                                 | 51 |
| 3.4 Erkenntnisse über Retrievaltests zur Suchmaschinenevaluierung.....        | 52 |
| <br>  |    |
| 4 Relevanzanalyse niedrig geranker Google-Suchergebnisse.....                 | 54 |
| 4.1 Studiendesign.....  | 54 |
| 4.1.1 Wahl der Suchmaschine .....   | 55 |
| 4.1.2 Verwendete Suchanfragen .....   | 55 |
| 4.1.3 Anzahl der Suchergebnisse.....  | 55 |
| 4.1.4 Die Juroren.....  | 56 |
| 4.1.5 Bewertung der Relevanz.....   | 56 |
| 4.1.6 Kennzahlen zur Datenauswertung .....                                    | 57 |
| 4.2 Umsetzung des Retrievaltests im RAT .....                                 | 60 |
| 4.2.1 Verwendung des RAT Backend zum Anlegen des Testrahmens.....             | 61 |
| 4.2.2 Erhebung der Webdokumente mit dem Suchmaschinen-Scraper .....           | 62 |
| 4.2.3 Testdurchführung mit dem RAT Frontend.....                              | 62 |
| 4.2.4 Aufbereitung der Daten mit dem RAT Auswertungsmodul.....                | 64 |
| 4.3 Ergebnisse der Studie.....  | 65 |
| 4.3.1 Häufigkeiten der Relevanzbewertung für alle Positionen .....            | 65 |
| 4.3.2 Häufigkeiten der Relevanzbewertung pro Position .....                   | 67 |
| 4.3.3 Häufigkeiten der Relevanzbewertung der Seite 1 und 6 im Vergleich. .... | 68 |
| 4.3.4 Häufigkeiten der Relevanzbewertung pro Suchanfrage.....                 | 70 |
| 4.3.5 Zusammenhang zwischen Position und Relevanz .....                       | 71 |
| 4.4 Ergebnisse der Datenanalyse .....   | 73 |

|  |     |
|--|-----|
| 4.4.1 Retrievaleffektivität der Suchmaschine bei den hinteren Positionen ... | 73  |
| 4.4.2 Beurteilung des Relevanzrankings .....                                 | 73  |
| 4.4.3 Zusammenhang zwischen Relevanz und Position .....                      | 74  |
| <br>   |     |
| 5 Fazit und Ausblick.....  | 75  |
| <br>   |     |
| Literatur.....   | 76  |
| Anhang.....  | 85  |
| Eidesstattliche Erklärung.....   | 117 |

## Abbildungsverzeichnis

|   |    |
|---|----|
| Abbildung 1: "Golden Triangle" & "F-Pattern" in der Heatmap (vgl. ENQUIRO 2005) ..... | 22 |
| Abbildung 2: Universal Search Ergebnisse in der Heatmap (vgl. ENQUIRO 2008) .....     | 24 |
| Abbildung 3: Die 5 Schritte zum Retrievaltest (vgl. LEWANDOWSKI 2012).....            | 43 |
| Abbildung 4: Darstellung des Userinterface anhand einer beispielhaften Suchaufgabe... | 64 |
| Abbildung 5: Histogramm zur Häufigkeitsverteilung für alle Relevanzwerte .....        | 66 |
| Abbildung 6: Durchschnittliche Relevanzbewertung pro Position .....                   | 67 |
| Abbildung 7: Median der Relevanzbewertung pro Position .....                          | 68 |
| Abbildung 8: Durchschnittliche Relevanzbewertung pro Position der Seite 1 und 6.....  | 69 |
| Abbildung 9: Median der Relevanzbewertung pro Position der Seite 1 und 6 .....        | 70 |
| Abbildung 10: Anzahl höherer Relevanzwerte bei Seite 1 und 6 pro Position .....       | 71 |

## Tabellenverzeichnis

|  |    |
|--|----|
| Tabelle 1: Erhebungsmethoden des Nutzerverhaltens in Suchmaschinen im Überblick .. | 19 |
| Tabelle 2: Vergleich der Retrievaltests hinsichtlich Methodik und Ergebnisse ..... | 53 |
| Tabelle 3: Häufigkeiten aller Positionen.....                                      | 65 |
| Tabelle 4: Korrelationsanalyse der Daten aus dem Retrievaltest .....               | 72 |
| Tabelle 5: Regressionsanalyse der Daten aus dem Retrievaltest .....                | 72 |
| Tabelle 6: R-Quadrat Test der Daten aus dem Retrievaltest.....                     | 72 |

## 1 Einleitung

Seit mehreren Jahren spielen Suchmaschinen bei der Online-Recherche eine entscheidende Rolle und gehören zu den meistgenutzten Internetanwendungen (vgl. VAN EIMEREN/FREES 2012, S. 369). Besonders der Marktführer Google wird von vielen Internetnutzern hierzulande als auch global als Suchwerkzeug intensiv benutzt (vgl. WEB-STATS.INFO 2013, WEBHITS.DE 2013, COMSCORE.COM 2013).

Durch die Eingabe einzelner Suchwörter über eine Benutzeroberfläche der Suchmaschine werden die Inhalte aus dem World Wide Web für den Internetnutzer recherchierbar gemacht, nachdem diese mittels Crawling erfasst worden sind (vgl. LEWANDOWSKI 2013, S. 495). Ziel der Suchmaschine ist das schnelle Finden und Bereitstellen von relevanten und zahlreichen Webdokumenten auf die Suchanfragen der Nutzer. Auf den Ergebnisseiten der Suchmaschine (search engine results pages; SERPS) werden organische Suchergebnisse, AdWords Anzeigen, Universal Search Ergebnisse aus anderen Kollektionen und Shortcuts bzw. Smart Answers in Form von direkten Fakteninformationen angezeigt (vgl. LEWANDOWSKI 2009, S. 204). Die organischen Suchergebnisse sind die eigentlichen Treffer aus dem Web-Index der Suchmaschine und stehen im Fokus der Informationssuche. Sie werden nach einer vom System bewerteten Relevanz geordnet, die aus Rankingfaktoren, wie u.a. Textstatistik, Popularität, Aktualität und Lokalität der Webdokumente ermittelt wird (vgl. LEWANDOWSKI 2013, S. 501). Die Suchergebnisse werden meist in Form einer Ergebnisliste präsentiert (vgl. LEWANDOWSKI 2009, S. 204), die häufig durch die Einstreuung von passenden Universal Search Ergebnissen (z.B. aus der Bild- oder Newssuche) oder Shortcuts (z.B. direkte Wetteranzeige oder Aktienkurse) angereichert wird (vgl. QUIRMBACH 2009, S. 235).

Websuchmaschinen werden von Information Professionals, wie Informationswissenschaftlern oder Rechercheuren, und hauptsächlich von Laiennutzern aus privaten oder beruflichen Gründen und meist ohne Vorkennt-

nisse oder bestimmte Strategien, verwendet und bedienen eine heterogene Nutzergruppe (vgl. LEWANDOWSKI 2005, S.34). Inhaltlich wird häufig nach Informationen zur Wirtschafts- und Arbeitswelt, nach Personen und Orten, IT- und Computerthemen, Unterhaltung und Freizeit sowie Informationen aus dem Bereich Gesundheit und Wissenschaft recherchiert (vgl. LEWANDOWSKI 2006, S.6).

Die eingegebenen Suchanfragen können nach den unterschiedlichen Suchmotivationen und Zielen der Nutzer in drei Typen klassifiziert werden: die navigations-, informations- und transaktionsorientierten Suchanfragen. Das Ziel der navigationsorientierten Suchanfrage ist das (Wieder-) Finden einer bestimmten Webseite, die dem Nutzer bereits bekannt oder deren Existenz vermutet wird. Dabei kann es sich beispielsweise um die Suche nach einer Unternehmenswebseite handeln. In der Regel gibt es auf eine navigationsorientierte Suchanfrage ein richtiges Ergebnis. Wird dieses gefunden, gilt das Informationsbedürfnis als befriedigt (vgl. BRODER 2002). Mit informationsorientierten Suchanfragen möchte sich der Nutzer zu einem bestimmten Thema informieren und meist mehrere Dokumente dazu recherchieren und lesen. In dem Sinne stellen informationsorientierte Suchanfragen den klassischen Anfragetyp im Information Retrieval dar. Das Informationsbedürfnis ist nicht durch ein einziges Ergebnis befriedigt, sondern durch das Erlangen der gesuchten Informationen (vgl. BRODER 2002). Mit transaktionsorientierten Suchanfragen sollen Webseiten aufgefunden werden, auf denen der Nutzer eine bestimmte (Trans-)Aktion durchführen kann. Dabei kann es sich zum Beispiel um einen Kauf eines Artikels oder einen Download einer Datei handeln. Wird eine Webseite gefunden, auf der die gewünschte Aktion ausführbar ist, gilt auch das Informationsbedürfnis des Nutzers als befriedigt (vgl. BRODER 2002). Eine Aussage, wie sich der Anteil der Suchanfragetypen an der Gesamtheit der Suchanfragen verteilt, lässt sich durch unterschiedliche Studienergebnisse nicht festlegen. Allerdings sollte eine erfolgreiche Beantwortung aller drei Anfragetypen gewährleistet sein, um das Informationsbedürfnis der Nutzer

zu befriedigen und eine Zufriedenheit über die Informationssuche mit der Suchmaschine zu erreichen (vgl. THUROW/MUSICA, 2009, S.29).

Die meisten Nutzer von Suchmaschinen investieren weder viel Zeit, Aufwand noch Mühe bei der Formulierung der Suchanfrage. Diese ist mit 1-2 Wörtern sowohl in deutscher (vgl. HÖCHSTÖTTER/KOCH 2009) als auch in englischer Sprache (vgl. JANSEN/SPINK 2006, S.254-56) recht kurz und inhaltlich häufig allgemein gehalten. Auch die Nutzung von Booleschen Operatoren oder der Erweiterten Suche findet nur in den seltensten Fällen statt (vgl. JANSEN/SPINK 2006, S.256). Die Nutzungskompetenz kann insgesamt als gering bezeichnet werden.

Bei der Sichtung und Auswahl der Suchergebnisse, meist bestehend aus Titel, Beschreibungstext und URL (vgl. LEWANDOWSKI 2009, S. 204), kann ebenso ein typisches Nutzerverhalten beobachtet werden. Zahlreiche Untersuchungen haben gezeigt, dass die meisten Nutzer die Treffer auf den Positionen 1 und 2 am intensivsten wahrnehmen, weil sie dort die am besten gerankten Suchergebnisse vermuten und auf die Relevanzbewertung durch die Suchmaschine vertrauen (vgl. KEANE/O'BRIEN/SMYTH 2008, S 4-6, CUTRELL/GUAN 2007, S.417-420, PAN et al. 2007, S.812-815). Die Aufmerksamkeit der Nutzer richtet sich dabei auf den oberen linken Bereich der SERP und nimmt mit zunehmender Rankingposition der Treffer ab (vgl. GRANKA/FEUSNER/LORIGO 2008, S.18, RODDEN/FU 2006, S.19, GOOGLE MARKTFORSCHUNG GERMANY 2007, S.8, LORIGO et al. 2008, S.1044). Sie sind also nicht bereit mehrere Dokumente bei der Suche zu sichten. Ähnlich verhält es sich mit der Klickhäufigkeit, die bereits nach dem Treffer auf Position 1 rapide abnimmt (vgl. RODDEN/FU 2007, S.19) und u.a. durch Faktoren wie Vertrautheit, Relevanzempfinden des gefundenen Webdokuments und dem Vorkommen des Suchkeywords im Suchergebnis beeinflusst wird (vgl. VAN GISBERGEN/VAN DER MOST/AELEN 2007, S.4). Die Betrachtungsdauer von durchschnittlich 1,1 Sekunden pro Suchergebnis spricht außerdem für ein grobes Scannen statt eines genaueren Lesens der Su-

chergbniselemente (vgl. VAN GISBERGEN/VAN DER MOST/AELEN 2007, S.4). Suchergebnisse jenseits der ersten beiden Positionen und vor allem jenseits der ersten SERP finden durch dieses Nutzerverhalten kaum Beachtung, ohne je von den Nutzern auf ihre Relevanz geprft worden zu sein.

### **1.1 Zweck und Abgrenzung der Arbeit**

Durch die hohe Bedeutung von Websuchmaschinen bei der Informationsbeschaffung und deren Ziel, trotz der meist trivialen und simplen Suchmethoden der Nutzer, aus der Vielzahl von Webdokumenten die relevantesten auffindbar und erreichbar zu machen, ist die Frage nach der Qualitt dieser bzw. ihrer Suchergebnisse von groem Interesse. Zu beantworten ist diese Frage im Rahmen der Suchmaschinenforschung, die sich in die Forschungsfelder Information-Retrieval-Technologie, Suchmaschinenqualitt, Information Research, Nutzerverhalten & Benutzerfhrung und Suchmaschinenkonomie unterteilen lsst (vgl. LEWANDOWSKI 2008, S.3).

Im Feld der Information-Retrieval-Technologie steht die Entwicklung von Web Information Retrieval Verfahren im Vordergrund, die in der Suchmaschinentechnologie zum Einsatz kommen. Dabei kann es sich um Vorgnge wie das Crawling oder das Ranking handeln. Auch Untersuchungen zum Einsatz semantischer Verfahren in der Websuche gehren zu diesem Feld der Suchmaschinenforschung (vgl. LEWANDOWSKI 2008, S.2). Ziel des Information Research ist die Optimierung des Vorgang der Informationssuche. Dazu soll zum einen der Umgang mit Recherche- Werkzeugen geschult, zum anderem die Recherchekompetenz der Nutzer ausgebildet werden. Forschungsbedarf gibt es auf diesem Gebiet vor allem nach der Validitt der Suchmaschinentreffer (vgl. LEWANDOWSKI 2008, S.5). Im Feld der Suchmaschinenkonomie wird der wirtschaftliche Aspekt im Bereich der Internetsuchmaschinen erforscht. Dabei werden der Markt und die Geschftsmodelle wie auch das Thema Suchmaschi-

nenmarketing analysiert (vgl. LEWANDOWSKI 2008, S.7). Beim Forschungsfeld Nutzerverhalten & Benutzerführung will man in der Suchmaschinenforschung die Vorgehensweisen, Wünsche, Bedürfnisse und Ziele verschiedener Nutzergruppen bei der Benutzung von Suchmaschinen ermitteln. Diese sind besonders für eine am Nutzer orientierte Weiterentwicklung der Suchdienste wichtig (vgl. LEWANDOWSKI 2008, S.6). Die Suchmaschinenqualität als Forschungsfeld beschäftigt sich mit der Leistungsfähigkeit von Suchmaschinen. Diese wird u.a. erforscht, um Suchmaschinen mit anderen Information Retrieval Systemen zur Informationssuche im Web, wie etwa Datenbanken, zu vergleichen und um Qualitätsunterschiede unter den verschiedenen Anbietern auszumachen (vgl. LEWANDOWSKI 2008 S.4). Die Qualität von Suchmaschinen lässt sich auf vier Untersuchungsebenen evaluieren: die Qualitätsmessung anhand des Suchmaschinenindex hinsichtlich Größe, Vollständigkeit und Aktualität, der Suchfunktionen bezogen auf ihren Umfang und Funktionsfähigkeit, der Nutzerfreundlichkeit in Verbindung mit der Usability des Dienstes und der Suchergebnisse hinsichtlich ihrer Relevanz und Einzigartigkeit, wobei letztere als Kernbereich zur Messung der Retrievaleffektivität angesehen wird (vgl. LEWANDOWSKI/HÖCHSTÖTTER 2008, S.14).

Diese Bachelorarbeit soll der Evaluierung der Suchmaschine Google dienen, indem anhand einer Relevanzanalyse niedrig gerankte Suchergebnisse auf informationsorientierte Suchanfragen auf ihre Relevanz überprüft werden. Anschließend sollen sie mit den Relevanzwerten der Suchergebnisse vorderer Positionen verglichen werden. So soll die Leistungsfähigkeit der Suchmaschine hinsichtlich ihrer Retrievaleffektivität, also der Fähigkeit auf eine Suchanfrage relevante Dokumente auszugeben (vgl. LEWANDOWSKI 2011, S.204), erstmalig bis hin zu den hinteren SERP überprüft werden. Auch das systemeigene Relevanzranking der Suchmaschine soll mit einer Relevanzbewertung durch die Nutzer kontrolliert werden.

Vor diesem Hintergrund ist die Arbeit in das Forschungsfeld der Suchmaschinenqualität einzuordnen, wobei als Messungsebene die Suchergebnisse herangezogen werden. Zusätzlich könnten Erkenntnisse über das Relevanzempfinden der Nutzer außerhalb der gewohnten und typischerweise von ihnen bewerteten Suchergebnispositionen gewonnen werden, sodass ein Erkenntnisfortschritt auch im Forschungsfeld des Nutzerverhaltens zu erwarten ist.

Die Fokussierung auf die Evaluation von Google-Suchergebnissen in dieser Arbeit ist auf die besondere Popularität dieser Suchmaschine und ihrer dominierenden Rolle am Suchmaschinenmarkt zurückzuführen. Die Begrenzung auf informationsorientierte Suchanfragen ist mit ihrer Eigenschaft zu begründen, das Informationsbedürfnis der Nutzer meist durch das Sichten mehrerer Webdokumente zu befriedigen. Auch die Tatsache, dass dieser Suchanfragetyp den Sinn des klassischen Information Retrieval, also dem Finden bestehender Information (LEWANDOWSKI 2005, S.33) am ehesten widerspiegelt, spricht für diese Abgrenzung. In diesem Fall ist die Frage nach der Relevanz der hinteren Suchergebnisse und der Zuverlässigkeit des systemeigenen Ranking von besonderem Interesse. Bei transaktions- und vor allem navigationsorientierten Suchanfragen sind diese Fragen weniger wichtig, da diese Suche meist durch das Finden eines einzelnen Webdokuments ihren Zweck erfüllt.

Ziel der Bachelorarbeit ist es folgende Aussagen über die Retrievaleffektivität, das systemeigene Relevanzranking der Suchmaschine und das Nutzerverhalten von Google-Usern zu bestätigen oder zu verwerfen:

- H1: Wenn bei informationsorientierten Suchanfragen Webdokumente aus den hinteren SERP für relevant bewertet werden, ist die Retrievaleffektivität der Suchmaschine auch bei den hinteren Trefferpositionen gegeben
- H2: Wenn die Relevanzbewertung der Suchergebnisse durch die Nutzer mit fallender Rankingposition ebenfalls sinkt, ist das Rele-

vanzranking der Suchmaschine zuverlässig und das Nutzerverhalten berechtigt

- H3: Wenn die Relevanzbewertung der Suchergebnisse durch die Nutzer bei fallender Rankingposition gleich bleibt, ist das Relevanzranking der Suchmaschine nicht notwendig und das Nutzerverhalten berechtigt
- H4: Wenn die Relevanzbewertung der Suchergebnisse durch die Nutzer mit fallender Rankingposition steigt, ist das Relevanzranking der Suchmaschine fehlerhaft und das Nutzerverhalten fragwürdig

Durch die Klärung dieser Hypothesen können gegeben falls Ansätze zur Optimierung der Leistungsfähigkeit der Suchmaschine diskutiert und Möglichkeiten und Motive zur Weiterentwicklung des Nutzerverhaltens erwo-gen werden.

## **1.2 Lösungsansatz und Methodik der Arbeit**

Da in dieser Arbeit hauptsächlich die Fähigkeit der Suchmaschine, auf Suchanfragen relevante Treffer auszugeben, überprüft werden soll, handelt es sich hierbei um eine systemorientierte Evaluation. Im Gegensatz zur benutzerorientierten Evaluation, bei der die Nutzer in den Vordergrund der Untersuchung rücken, wird bei dieser Evaluierungsform der Fokus auf die systembezogenen Funktionen der Suchmaschine gesetzt (vgl. TURPIN/SCHOLER 2006, S.3).

Als Methode zur Durchführung der Relevanzanalyse soll ein Retrievaltest als Verfahren aus dem klassischen Information Retrieval, angepasst auf die Merkmale der Websuche (vgl. LEWANDOWSKI 2005, S.138), durchgeführt werden. Dabei werden standardmäßig eine bestimmte Menge an zuvor erhobenen realen Suchanfragen an eine Suchmaschine gestellt, die zurückgegebenen Ergebnisse anonymisiert und gemischt und auf eine bestimmte Trefferposition beschränkt (Cut Off –Wert, häufig bis Position 10) einer Gruppe von Juroren zur Relevanzbewertung vorgelegt. Die Bewertung erfolgt dabei aufgrund der persönlich empfundenen Relevanz und

stellt somit eine subjektive, vom Wissen und der Reflexion der Juroren abhängigen Einschätzung dar (vgl. QUIRMBACH 2012, S.131). Anschließend werden die Ergebnisse den Trefferplätzen zugeordnet, sodass eine Auswertung meist anhand der Precision, also der Fähigkeit der Suchmaschine, nur relevante Treffer auszugeben, stattfinden und Aussagen über die Qualität der Suchmaschine bzw. ihrer Suchergebnisse gemacht werden können (vgl. LEWANDOWSKI 2011, S.206).

Eine Alternative zum Retrievaltest als Mittel zur systemorientierten Evaluation der Suchergebnisqualität bietet die Klickdatenanalyse. Dabei werden Logfiles ausgewertet, die Informationen darüber bieten welcher Nutzer (an IP-Adresse, Login, Cookies etc. identifizierbar) nach welchen Suchanfragen welche Suchergebnisse angeklickt hat (vgl. HÖCHSTÖTTER 2009, S.179). So können die Anzahl der Klicks und die Verweildauer des Nutzers auf einem Suchergebnis gemessen und Rückschlüsse auf die Relevanz der Treffer gezogen werden.

Obwohl die Klickdatenanalyse weniger aufwendig ist als der Retrievaltest, der in seiner Konzeption und Durchführung etwa durch die Erhebung realer Suchanfragen und das Akquirieren der Juroren mehr Zeit und Vorbereitung kostet, wurde in dieser Arbeit trotzdem letzterer als Methode zur Relevanzanalyse niedrig geranker Google-Suchergebnisse gewählt. Durch den Retrievaltest wird die direkte Erhebung von Relevanzurteilen ermöglicht. Besonders vor dem Hintergrund, Webdokumente jenseits der ersten SERP auf ihre Relevanz zu untersuchen, wäre eine Klickdatenanalyse weniger sinnvoll, da aufgrund des gängigen Nutzerverhaltens kaum von Klickraten auf Suchergebnissen dieser Positionen auszugehen ist.

Zur Umsetzung des Retrievaltests hinsichtlich des Testdesign und der Erfassung, Aufbereitung und Auswertung der Daten soll das Relevance Assessment Tool (RAT) verwendet werden. Dabei handelt es sich um ein Onlinewerkzeug, welches den Aufwand, der im Zusammenhang mit der Untersuchung von Suchergebnissen als Testobjekte einherging, reduzie-

ren soll (vgl. SÜNKLER 2012). Das RAT besteht aus vier Komponenten: den Suchmaschinen-Scraper, der die Suchergebnisse von der Suchmaschine kopiert und speichert, dem Backend zum Anlegen und Definition des Testrahmens, dem Frontend zur eigentlichen Datenerhebung der Relevanzbewertungen durch die Juroren und dem Auswertungsmodul zur Aufbereitung und Analyse der Daten (vgl. SÜNKLER 2012, S.65). Aufgrund dieser Eigenschaften reduziert das RAT den mit Retrievaltests verbundenen Aufwand und soll im Rahmen der Bachelorarbeit zur effizienteren Umsetzung der angestrebten Untersuchung zum Einsatz kommen.

### **1.3 Struktur der Arbeit**

Die Arbeit setzt sich insgesamt aus vier Teilen zusammen, die jeweils in mehrere Kapitel gegliedert sind. Im ersten Teil wird zunächst die Forschung im Bereich des Nutzerverhaltens in Suchmaschinen näher erläutert. Dabei soll der Fokus auf der Beantwortung der Frage liegen, inwiefern die Position der Suchergebnisse im Relevanzranking der Suchmaschine Einfluss auf die Wahrnehmung und Evaluierung der Treffer durch die Nutzer hat. Es werden die Methoden mit ihren Vor- und Nachteilen zur Klärung dieser Frage diskutiert, Studien zu diesem Thema vorgestellt und die Erkenntnisse aus diesem Forschungsbereich zusammengefasst. Im zweiten Teil wird die systemorientierte Evaluation von Suchmaschinen thematisiert, wobei der Retrievaltest als Methode im Vordergrund stehen soll. Es werden der Begriff der Relevanz diskutiert, Standards und Normen vorgestellt, die sich in der langen Tradition der Evaluierung von Information Retrieval Systemen etabliert haben, einige bereits durchgeführte Retrievaltests erläutert und die Ergebnisse daraus resümiert. Im dritten Teil soll die Studie zur Relevanzanalyse niedrig geranker Google-Suchergebnisse auf informationsorientierte Suchanfragen vorgestellt werden. Dabei werden zunächst das Studiendesign erläutert, die Durchführung beschrieben und die Ergebnisse der Studie vorgelegt werden. Anschließend sollen diese auf die Fragen nach der Retrievaleffektivität von

Google bis hin zu den hinteren SERP und der Zuverlässigkeit des systemeigenen Relevanzrankings interpretiert und diskutiert werden.

## **2 Selektions- & Navigationsverhalten der Nutzer auf SERP**

Ein Verständnis für die Verhaltensweisen, Ziele und Motive der Suchmaschinennutzer ist in der Suchmaschinenforschung von großer Bedeutung, auch hinsichtlich der Suchmaschinenqualität. Schließlich lassen sich durch die Erkenntnisse aus dem Nutzerverhalten direkte Maßnahmen zur Optimierung und Verbesserung dieser ableiten. Auch die Problematik, dass in der Suchmaschinennutzung Suchergebnisse jenseits der ersten beiden Treffer kaum wahrgenommen werden und so die dort aufgeführten Webdokumente kein Relevanzurteil durch die Nutzer erhalten, ist u.a. der Grund für die Evaluierung des Rankingverfahrens und der Retrievaleffektivität der Suchmaschine bis hin zu den hinteren SERP in dieser Arbeit. Die Relevanzbewertung niedrig geranker Google-Suchergebnisse durch Juroren dient deshalb nicht nur als Evaluationsmethode sondern soll auch die Frage beantworten, ob das Selektions- und Navigationsverhalten der Suchmaschinennutzer gerechtfertigt ist. Wie das Selektions- und Navigationsverhalten erforscht werden kann, zu welchen Erkenntnissen Studien auf diesem Gebiet gekommen und wie sie zu werten sind, steht im Vordergrund dieses Kapitels.

### **2.1 Analysemethoden des Selektions- & Navigationsverhalten auf SERP**

Das Nutzerverhalten, auch bezogen auf die Selektion und Navigation auf den SERP einer Suchmaschine, kann als recht gut erforschter Bereich angesehen werden (vgl. LEWANDOWSKI 2008, S.6). Dabei bieten sich zur Erhebung mehrere Methoden an. Grundsätzlich kann man diese in überwachte und unüberwachte unterteilen. Während bei überwachten Methoden dem Nutzer klar ist, dass seine Handlungen beobachtet, erhoben und analysiert werden, erfolgen dieselben Prozesse bei unüberwachten Methoden meist automatisiert und ohne Kenntnis des Nutzers. Beide Er-

hebungsmethoden können auch zusammen angewendet werden, wobei eine Methode zusammengefasst als überwacht bezeichnet werden muss, sobald nur ein Bestandteil der Datenerhebung von dem Nutzer als solcher erkennbar ist (vgl. HÖCHSTÖTTER 2009, S.176).

### **2.1.1 Überwachte Erhebungsmethoden**

Sobald der Versuchsperson bewusst ist, dass ihr Verhalten in der Suchmaschinennutzung zu Erhebungszwecken beaufsichtigt und anschließend analysiert wird, spricht man von einer überwachten Erhebungsmethode (vgl. HÖCHSTÖTTER 2009, S.176). Diese bietet mehrere Vorgehensweisen, die in den nachfolgenden Kapiteln mit ihren Vor- und Nachteilen vorgestellt werden.

#### **2.1.1.1 Fragebögen**

Zur überwachten Methode gehört die Erhebung durch Fragebögen in papiergebundener Form mit einem persönlichen Interview oder als elektronische Onlineumfrage. Dabei können Nutzer direkt nach ihren Such bzw. Selektions- und Navigationsverhalten befragt werden. Als besonders praktisch bei der Erhebung von internetbezogenen Daten ist die Onlinebefragung anzusehen. Bei ihr liegen die Daten bereits in elektronischer Form vor, sodass für eine computerbasierte statistische Auswertung dieser, kein zusätzlicher Aufwand durch Übertragen oder Einscannen entsteht (vgl. HÖCHSTÖTTER 2009, S.176). Bei der Gestaltung des Fragebogens ist zu unterscheiden zwischen geschlossenen Fragen, bei denen vordefinierte Antworten möglich sind und offenen, bei denen der Befragte in der Formulierung seiner Antwort komplett frei ist. Eine besondere Bedeutung haben auch (meist 5-stufig) skalierte Fragen, die zur Berechnung von differenzierten Varianz- und Mittelwerten dienen und Summenfragen, bei denen der Befragte beispielsweise 100 Punkte auf seine Antworten verteilen kann, um die Gewichtung dieser zu verdeutlichen. Dies kann auch durch eine Reihenfolge, die vom Befragten zu vergeben ist, verdeutlicht werden.

Insgesamt ist auf eine einfache und leicht verständliche Formulierung zu achten (vgl. HÖCHSTÖTTER 2009, S.176).

Der Fragebogen als Erhebungsmethode für Nutzerdaten bietet einige Vorteile. Zum einen können so Hintergrundinformationen, wie etwa demographische Angaben der Teilnehmer erfragt werden. So können Nutzergruppen identifiziert und miteinander verglichen werden. Die elektronische Verfügbarkeit der Daten und der Prozess der Zwischenspeicherung der Antworten bei abgebrochenen Vorgängen ist ebenfalls als positiv anzusehen (vgl. HÖCHSTÖTTER 2009, S. 176; HÖCHSTÖTTER 2007, S.137).

Ein Nachteil stellt die Tatsache dar, dass viele Teilnehmer in einer beobachteten Situation dazu neigen sich in ihren Kenntnissen und Fähigkeiten besser darstellen zu wollen, als in der Realität der Fall ist. Auch sind nicht alle Fragen nach der Nutzung von Suchmaschinen dafür geeignet, in dieser Erhebungsmethode gestellt zu werden. Die durchschnittliche Betrachtungsdauer der Suchergebnistreiber etwa kann kaum durch Fragebögen erforscht werden, weil dies für die meisten Befragten ein allzu unbewusster Vorgang ist. Auch ist das Thema der Repräsentativität der Umfrageergebnisse durch die heterogene Nutzergruppe in Suchmaschinen problematisch. Bei der Wahl der Teilnehmer muss deshalb auf eine möglichst gemischte Zielgruppe geachtet werden, oder es werden bewusst Daten einer speziellen Zielgruppe, z.B. begrenzt auf die Wahl einer Suchmaschine, erhoben (vgl. HÖCHSTÖTTER 2009, S.176; HÖCHSTÖTTER 2007, S.137).

#### **2.1.1.2 Laborexperimente**

Eine weitere überwachte Methode stellt die Erhebung von Nutzerdaten durch (Computer-)Laborexperimente dar. Dabei wird die Testperson in seiner Interaktion mit der Suchmaschine am Computer direkt beobachtet. Die Laborexperimente werden in einem dafür ausgestatteten Raum durchgeführt, wobei in diesem Fall die Verfügbarkeit eines Computers mit Internetverbindung besonders wichtig ist. Für die Beobachtung und Auf-

zeichnung des Nutzerverhaltens der Testperson ist ein Interviewer anwesend (vgl. HÖCHSTÖTTER 2009, S.177). Bei der Untersuchung des Nutzerverhaltens in Suchmaschinen werden den Testpersonen meist Suchaufgaben gestellt, die direkt gelöst werden sollen. Die Vorgehensweise der Teilnehmer, von der Formulierung der Suchanfrage, über die Auswahl der Suchergebnisse bis hin zum Verhalten auf dem angeklickten Webdokument und darüber hinaus wird vom Interviewer detailliert protokolliert und nachvollzogen. Häufig wird das Laborexperiment durch einen Fragebogen ergänzt, durch den am Ende der Erhebung weitere Hintergrundinformationen, etwa zur Internetnutzung allgemein oder die demographischen Angaben der Testpersonen erhoben werden. (vgl. HÖCHSTÖTTER 2009, S.177).

Das Laborexperiment zur Erhebung des Nutzerverhaltens in Suchmaschinen, auch in Bezug auf das Selektions- und Navigationsverhalten, bietet den Vorteil, dass die Vorgehensweisen der Testpersonen direkt und detailliert beobachtet und protokolliert werden können. Auch können Hintergrundinformationen etwa zur Begründung des Verhaltens hinzugezogen werden, da Rechenschritte der Probanden einzeln erklärt und nachvollzogen werden (vgl. HÖCHSTÖTTER 2009, S.177).

Nachteilig sind allerdings die hohen Kosten der Methode, die durch die Ausstattung der Labore verursacht werden und so zu einer limitierten Anzahl an Arbeitsplätzen und Räumlichkeiten führen. Dadurch und durch den hohen Zeitaufwand fällt die Teilnehmerzahl in solchen Experimenten meist gering aus, sodass nicht von einer repräsentativen Erhebung gesprochen werden kann. Auch kann von einem veränderten Verhalten der Testpersonen ausgegangen werden, die durch die Anwesenheit der Interviewer und das Experimentszenario beeinflussbar sind. Zudem spiegeln die gestellten Suchanfragen häufig nicht das reale Informationsbedürfnis wieder, welches bei der Internetrecherche aufkommt (vgl. HÖCHSTÖTTER 2009, S.177; HÖCHSTÖTTER 2007, S.137).

### 2.1.1.3 Blickaufzeichnung / Eye Tracking

Das Aufzeichnen der Augen- und Blickbewegung einer Person gehört ebenfalls zur überwachten Erhebungsmethode und ist als Laboruntersuchung zu kategorisieren. Diese wird auch als Eye Tracking bezeichnet, wobei zwischen Remote Eye Tracking, bei dem sich das Aufzeichnungsgerät vor der Testperson befindet, und Head-mounted Eye Tracking, bei dem das Gerät auf dem Kopf des Probanden befestigt wird, zu unterscheiden ist. Erstere kommt bei der Untersuchung des Nutzerverhaltens im Web und speziell bei der Suchmaschinennutzung häufiger zum Einsatz, da die Mobilität, die beim Head-mounted Eye Tracking gewährleistet wird, nicht zwingend notwendig ist und so Zeit und Aufwand zum Montieren des Geräts gespart wird. Für das Erforschen des Selektions- und Navigationsverhaltens der Nutzer auf SERP ist diese Methode besonders hervorzuheben, da durch die Blickaufzeichnung der Probanden ihre Wahrnehmungsintensität und –Länge direkt erhoben werden kann (vgl. HÖCHSTÖTTER 2009, S.178). Fokussiert eine Testperson einen bestimmten Punkt auf der SERP, wird dieser auf einer Heatmap visualisiert und stellt eine sogenannte „Area of Interests“ (AOI) dar. Die am häufigsten bzw. am längsten betrachteten Bereiche werden rot dargestellt, während vernachlässigte Bereiche blau markiert werden. Die so entstehende Karte gibt Auskunft über aufmerksamkeitsstarke (heiße) und aufmerksamkeitschwache (kalte) Bereiche auf der SERP. Neben den Blicken können nach demselben Prinzip auch Mausbewegungen und Klicks als Aufmerksamkeitsbewertungen der Probanden definiert und aufgezeichnet werden. So wird erforscht, welche Bereiche der SERP den Nutzer interessieren und welche Trefferpositionen die größte Aufmerksamkeit erzeugen (vgl. HÖCHSTÖTTER 2009, S.178).

Die Methode der Blickaufzeichnung bietet den entscheidenden Vorteil, dass unbewusste Bewegungen und Blicke aufgezeichnet und visualisiert werden können, die die Testpersonen selbst nicht reflektieren und erläu-

tern können. So kann das Evaluationsverhalten der Nutzer direkt beobachtet und erhoben werden (vgl. HÖCHSTÖTTER 2009, S.178).

Es entsteht allerdings die Problematik, dass die Probanden wissen, dass sie beobachtet werden. Da Blicke allerdings Reflexe sind, die nicht bewusst gesteuert oder geändert werden können, ist ein abweichendes Verhalten vom tatsächlichen Alltagsszenario kaum zu erwarten. Der hohe Kosten- und Zeitaufwand bleibt dennoch bestehen (vgl. HÖCHSTÖTTER 2009, S.178).

#### **2.1.1.4 Online Panels**

Bei einer Untersuchung mittels Online Panels werden die gleichen Probanden bezüglich stets gleicher Merkmale und Untersuchungsfragen über einen längeren Zeitraum wiederholt untersucht. Dabei wird eine Software auf den Computern der Teilnehmer installiert, die die Internetnutzung protokolliert und in Click-Stream-Daten speichert. Zusätzlich können Onlineumfragen die Paneluntersuchung ergänzen (vgl. HÖCHSTÖTTER 2009, S.179).

Die Untersuchungsmethode bietet Hintergrundinformationen zu demographischen Merkmalen der Testpersonen, was als Vorteil einzustufen ist. Auch die Tatsache, dass Daten kontinuierlich und über einen längeren Zeitraum erhoben werden um Häufigkeiten und Veränderungen in der Suchmaschinennutzung festzustellen, spricht für diese Methode (vgl. HÖCHSTÖTTER 2009, S.179).

Als Nachteil bei der Untersuchungsmethode ist die sogenannte Panelsterblichkeit anzusehen, die auftritt, wenn Testpersonen mit der aktiven Teilnahme an der Erhebung aufhören. Auch der Paneleffekt kann häufig beobachtet werden, also eine veränderte in Internet- und Suchmaschinennutzung bedingt durch die Kenntnis der Testpersonen über die Beobachtung. Die Frage nach der Repräsentativität der erhobenen Daten

kann ebenfalls als problematisch angesehen werden (vgl. HÖCHSTÖTTER 2009, S.179).

## **2.1.2 Unüberwachte Erhebungsmethoden**

Im Gegensatz zu überwachten Erhebungsmethoden werden die Teilnehmer bei dieser Vorgehensweise zur Datenerhebung in ihrem Nutzerverhalten beobachtet, ohne davon in Kenntnis gesetzt zu sein. Die Protokollierung und Erhebung der Daten findet automatisiert und ohne Anwesenheit eines Interviewers statt und wird elektronisch gespeichert. Auch die unüberwachte Erhebungsmethode bietet einige Vorgehensweisen, die im Folgenden vorgestellt werden (vgl. HÖCHSTÖTTER 2009, S.180).

### **2.1.2.1 Logfileanalyse**

Zu einer sehr häufig verwendeten unüberwachten Methode zur Untersuchung des Nutzerverhaltens in Suchmaschinen gehört die Logfileanalyse. Eine Logfile bzw. Logdatei eines Webserver, meist im Format NCSA (Common Logfile Format) protokolliert alle Prozesse, Zugriffe, Einstellungen und Aktivitäten bei der Internetnutzung und wird automatisch zur Datenspeicherung erstellt. In Hinblick auf das Nutzerverhalten in Suchmaschinen können so Suchanfragen und -Vorgänge erhoben und durch die Identifikation anhand der IP Adresse oder Cookies einzelnen Nutzern zugeordnet werden, die durch diese Daten wiedererkannt werden, allerdings namentlich anonym bleiben. Von der Wahl der Suchmaschine, über die Formulierung der Suchanfrage, Benutzung von Operatoren und Hilfen, Verweildauer auf der SERP, Auswahl des Suchergebnisses und Verhalten auf dem angeklickten Webdokument, bis hin zur Rückkehr auf die SERP oder Neuformulierung der Suchanfrage, kann das Verhalten der Nutzer erhoben werden. Auch allgemeine Aussagen, etwa über die Anzahl der gestellten Suchanfragen einer Person in einem gewissen Zeitraum, der Dauer einer Suchsession oder über die häufigsten und populärsten Suchanfragen pro Monat können durch die Logfileanalyse gemacht werden (vgl. HÖCHSTÖTTER 2009, S.180).

Die Tatsache, dass eine besonders große Datenmenge und eine breite Nutzergruppe vorhanden sind und kaum zusätzlicher Aufwand für die Erhebung betrieben werden muss, stellt einen entscheidenden Vorteil dieser Methode dar. Auch kann das Verhalten der Nutzer wie es real im Alltag stattfindet, beobachtet werden ohne durch Testszenarien oder Interviewer beeinflusst zu werden. Es kann von objektiven und repräsentativen Ergebnissen gesprochen werden (vgl. HÖCHSTÖTTER 2009, S.180; HÖCHSTÖTTER 2007, S.137).

Der ethische Aspekt, dass Daten von häufig auch sehr persönlicher Natur ohne Zustimmung der Probanden erhoben werden, ist ein Nachteil der Logfileanalyse. Sobald ein Nutzer nach seinen eigenen Namen sucht, wie es im Phänomen des Egosurfings bzw. Vanity Search der Fall ist (vgl. NICOLAI et al. 2009, S.1), kann auch von keiner namentlichen Anonymität der Nutzer gesprochen werden. Auch ist nicht eindeutig festzustellen, wann eine Suche beendet wird. Deshalb wird meist eine zeitliche Begrenzung, etwa wenn der Nutzer innerhalb von 10 Minuten nicht vom Webdokument zur Suchmaschine zurückkehrt, dafür definiert. Die nächste Suchanfrage des Nutzers zählt dann als Eröffnung einer neuen Suchsitzung. Bei der Logfileanalyse fehlen außerdem Angaben zur Demographie des Nutzers, sodass keine Bildung von Nutzergruppen oder Vergleichen möglich ist. Ausschlaggebend ist allerdings der Punkt, dass Logfiles meist nicht zur freien Verwendung freigegeben werden und deren Benutzung von Kooperationen mit den Suchmaschinenanbietern abhängt. (vgl. HÖCHSTÖTTER 2009, S.180; HÖCHSTÖTTER 2007, S.137).

### **2.1.2.2 Livetickeranalyse**

Eine weitere Möglichkeit zur Datenerhebung stellt die Analyse von Livetikern dar, die von einigen Suchmaschinen angeboten werden. Dabei können die zeitgleich von realen Nutzern eingegebenen Suchanfragen gesichtet werden, die regelmäßig und meist in Echtzeit aktualisiert werden. Im Rahmen der Erforschung des Nutzerverhaltens werden diese Daten

durch ein Programm automatisch abgerufen, aktualisiert, erhoben und im Anschluss analysiert (vgl. HÖCHSTÖTTER 2009, S.181).

Diese Methode bietet den Vorteil, dass eine große Datenbasis meist ohne viel Zeitaufwand und kostengünstig über einen längeren Zeitraum erhoben werden kann, da solche Daten in der Regel frei zugänglich sind. Auch entsteht keine Situation, durch die Nutzer sich beobachtet fühlen und ihr Verhalten gegeben falls ändern könnten. Diese bleiben bei dieser Methode anonym (vgl. HÖCHSTÖTTER 2009, S.181).

Allerdings werden weder Hintergrundinformationen zum Nutzer, noch Daten erhoben, die über die Formulierung der Suchanfrage hinausgehen. Auch werden Suchanfragen nach bedenklichen Themen häufig von Livetickern ausgeschlossen, was zu Verzerrungen in der Datenauswertung führen könnte. Zusätzlich haben viele Suchmaschinenanbieter, darunter auch web.de und icerocket.com ihre Livetickerfunktionen eingestellt, sodass die Methode nicht zum Einsatz kommen kann (vgl. HÖCHSTÖTTER 2009, S.181).

### **2.1.3 Methodenüberblick- & Zusammenfassung**

Der Vergleich der verschiedenen Erhebungsmethoden, in Tabelle 1 zusammengefasst, zeigt, dass man bei den unbeobachteten Methoden von repräsentativeren Ergebnissen ausgehen kann. Zum einen werden meist größere Datenmengen erhoben und das Verhalten der Probanden entspricht durch die unbeobachtete Situation eher der Realität. Die beobachteten Methoden bieten allerdings den Vorteil, dass Hintergrundinformationen zum Nutzer erhoben werden, wodurch die Auswertungen detaillierter und differenzierter ausfallen.

Auch kann festgestellt werden, dass nicht jede Methode für die Erfassung des Evaluationsverhaltens durch die Nutzer bei der Trefferwahl geeignet ist.

|                    | <b>Methode</b>    | <b>Vorteile</b>   | <b>Nachteile</b>   |
|--------------------|-------------------|---|--|
| <b>überwacht</b>   | Fragebögen        | <ul style="list-style-type: none"> <li>• Demographie- &amp; Detailfragen</li> </ul>   | <ul style="list-style-type: none"> <li>• Beeinflussung</li> <li>• wenig repräsentativ</li> </ul>   |
|                    | Laborexperiment   | <ul style="list-style-type: none"> <li>• Demographie- &amp; Detailfragen</li> <li>• direkte Beobachtung</li> </ul>  | <ul style="list-style-type: none"> <li>• Beeinflussung</li> <li>• Kosten- und Zeitaufwand</li> <li>• geringe Teilnehmerzahl</li> <li>• wenig repräsentativ</li> </ul>              |
|                    | Eye Tracking      | <ul style="list-style-type: none"> <li>• Demographie- &amp; Detailfragen</li> <li>• direkte Beobachtung der Reflexe und unbewusster Verhaltensweisen</li> </ul>               | <ul style="list-style-type: none"> <li>• Beeinflussung</li> <li>• Kosten- und Zeitaufwand</li> <li>• geringe Teilnehmerzahl</li> <li>• wenig repräsentativ</li> </ul>              |
|                    | Online Panels     | <ul style="list-style-type: none"> <li>• Demographie- &amp; Detailfragen</li> <li>• direkte Beobachtung</li> <li>• kontinuierliche Erhebung</li> </ul>                        | <ul style="list-style-type: none"> <li>• Beeinflussung</li> <li>• Kosten- und Zeitaufwand</li> <li>• geringe Teilnehmerzahl</li> <li>• Panelsterblichkeit &amp; -Effekt</li> </ul> |
| <b>unüberwacht</b> | Logfileanalyse    | <ul style="list-style-type: none"> <li>• keine Beeinflussung</li> <li>• kontinuierliche Erhebung</li> <li>• Datenmenge &amp; Nutzergruppe</li> <li>• repräsentativ</li> </ul> | <ul style="list-style-type: none"> <li>• keine Demographie- und Detailfragen</li> <li>• Sessions-Problematik</li> <li>• Abhängigkeit von Suchmaschinenbetreiber</li> </ul>         |
|                    | Livetickeranalyse | <ul style="list-style-type: none"> <li>• keine Beeinflussung</li> <li>• kontinuierliche Erhebung</li> <li>• Datenmenge &amp; Nutzergruppe</li> <li>• repräsentativ</li> </ul> | <ul style="list-style-type: none"> <li>• keine Demographie- und Detailfragen</li> <li>• Filter bedenklicher Inhalte</li> <li>• viele eingestellt</li> </ul>                        |

**Tabelle 1: Erhebungsmethoden des Nutzerverhaltens in Suchmaschinen im Überblick**

Zum Erforschen des Selektions- und Navigationsverhalten der Nutzer auf der SERP ist die Blickaufzeichnung als Methode besonders effektiv, da durch das Eye Tracking Verfahren die reflexartige und unbewusste Wahrnehmung der SERP direkt erhoben, visualisiert und anschließend interpretiert wird. Von der Wahrnehmung auf der SERP kann so auf die Evaluierung der Suchergebnisse durch die Nutzer schließen. Aus diesem Grund werden im nachfolgenden Kapitel ausschließlich Eye Tracking-Studien vorgestellt. Mit ihnen soll die Frage beantwortet werden, wie die Nutzer bei der Relevanzbewertung und Auswahl der Suchtreffer vorgehen und welche Rolle dabei ihre Position in der Ergebnisliste spielt.

## **2.2 Stand der Forschung zum Selektions- & Navigationsverhalten auf SERP**

Die Eye Tracking Methode wird zur Untersuchung des Evaluationsverhaltens der Nutzer auf Suchergebnisseiten häufig verwendet. Grundsätzlich kann man die durchgeführten Studien in jene einteilen, die als Untersuchungsobjekt unveränderte SERP verwendeten und die, bei denen den Testpersonen manipulierte Ergebnisseiten vorgelegt werden. Im Folgenden werden einige Eye Tracking Studien beider Kategorien vorgestellt. Das Studiendesign, die Durchführung und die Ergebnisse sollen erläutert und anschließend interpretiert und ihr Wert in Hinblick auf die Forschungsfragen und Hypothesen dieser Arbeit beurteilt werden.

### **2.2.1 Studien zu unveränderten SERP**

Um Erkenntnisse darüber zu gewinnen, wie das tatsächliche Nutzerverhalten bezogen auf die Selektion und Navigation der Suchergebnisse einer SERP ist, werden Probanden in Eye Tracking Studien meist Suchaufgaben gegeben, die sie mithilfe einer Suchmaschine lösen sollen. In der Regel wird dabei die Suchanfrage von den Probanden selbst formuliert und an die Suchmaschine geschickt. Die daraufhin ausgegebene SERP, nicht manipuliert und im Originalzustand, und ihre Wahrnehmung durch den Probanden liegen im Fokus der Untersuchungen.

#### **2.2.1.1 Die Studie von Granka et al.**

Die Eyetracking-Methode zur Untersuchung des Nutzererhaltens in Suchmaschinen wurde erstmals von Granka et al. im Jahr 2004 mit 36 Probanden angewendet. Diese bekamen jeweils 10 Suchaufgaben zu gemischten Themen, von denen 5 navigationsorientiert und 5 informationsorientiert waren. Diese sollten mithilfe der Suchmaschine Google gelöst werden. Insgesamt wurden so 397 Suchanfragen abgeschickt (vgl. GRANKA et al. 2004, S.1).

Als Zentrale Erkenntnisse der Arbeit kam heraus, dass die Probanden die ersten beiden Treffer am intensivsten wahrgenommen haben, wobei der Treffer auf der ersten Position bei weitem am häufigsten geklickt wurde. Zudem konnte festgestellt werden, dass die Wahrnehmung der Treffer ab Position 5 besonders gering war, was damit erklärt wurde, dass sich diese Suchergebnisse „below the fold“, also außerhalb des sofort sichtbaren Bereich des Bildschirms befanden. Als drittes Ergebnis wurde beobachtet, dass die Testpersonen dazu neigten die Suchergebnisse der Reihe nach von oben nach unten zu sichten. Es konnte somit festgestellt werden, dass die Evaluierung der Nutzer stark durch das Ranking der Treffer beeinflusst wird (vgl. GRANKA et al. 2004, S.2).

#### **2.2.1.2 Die Enquiro-Studien von Hotchkiss et al.**

Eine Studie, die ebenfalls die hohe Bedeutung der Position der Treffer bei der Wahrnehmung der SERP durch die Nutzer erkannte, war die Enquiro-Studie von Hotchkiss et al. aus dem Jahr 2005. 48 Testpersonen wurden für die Studie nach der Remote Eye Tracking Methode akquiriert. Ihnen wurden jeweils 5 Suchaufgaben vorgegeben, die mithilfe der Suchmaschine Google gelöst werden sollten. Nach der Eingabe der Suchbegriffe durch die Testpersonen erfolgte die Blickaufzeichnung durch den Eye Tracker. Die ermittelte Aufmerksamkeitsverteilung auf den SERP wurde in Form einer Heatmap visualisiert (vgl. ENQUIRO 2005, S.13).

Als Ergebnis der Studie kam heraus, dass sich die Blicke der Nutzer beim erstmaligen Erscheinen der SERP in der linken oberen Ecke auf die ersten beiden organischen Suchergebnisse konzentrieren und dann nach unten und nach rechts folgen. In den Heatmaps konnte dieses Verhalten durch einen hervorgehobenen dreieckigen Bereich identifiziert werden. Deshalb wurde vom „Golden Triangle“ Effekt gesprochen. Die Blickfolge zum unteren Bereich der SERP und nach rechts entlang der Suchergebnistreffer wurde als „F-shaped pattern“ bezeichnet (vgl. Abbildung 1). Besonders der Bereich der SERP „below the fold“ wurde von den meisten Testpersonen weniger beachtet. Während die ersten drei Positionen von

allen Probanden fokussiert wurden, waren es bis zur Position 5 nur noch 60%, und bis zur Position 10 lediglich 20% der Testpersonen. So wurde herausgestellt, dass die Position der Suchergebnisse eine entscheidende Rolle für die Wahrnehmung und anschließende Selektion dieser spielt und die meisten Nutzer sich auf die Relevanzbewertung der Suchmaschine verlassen (vgl. ENQUIRO 2005, S.13).

Zu ähnlichen Erkenntnissen kam eine zwei Jahre später von Google selbst veröffentlichte Studie, bei der in den Heatmaps ebenso das Dreiecks- bzw. F-Muster identifiziert wurde. Diese Fokussierung der Probanden wurde mit dem „Google Effekt“ erklärt, durch den die Nutzer es gewohnt seien, die relevantesten Webdokumente durch das Ranking der Suchmaschine auf den ersten beiden Positionen zu finden (GOOGLE MARKTFORSCHUNG GERMANY 2007, S.7).

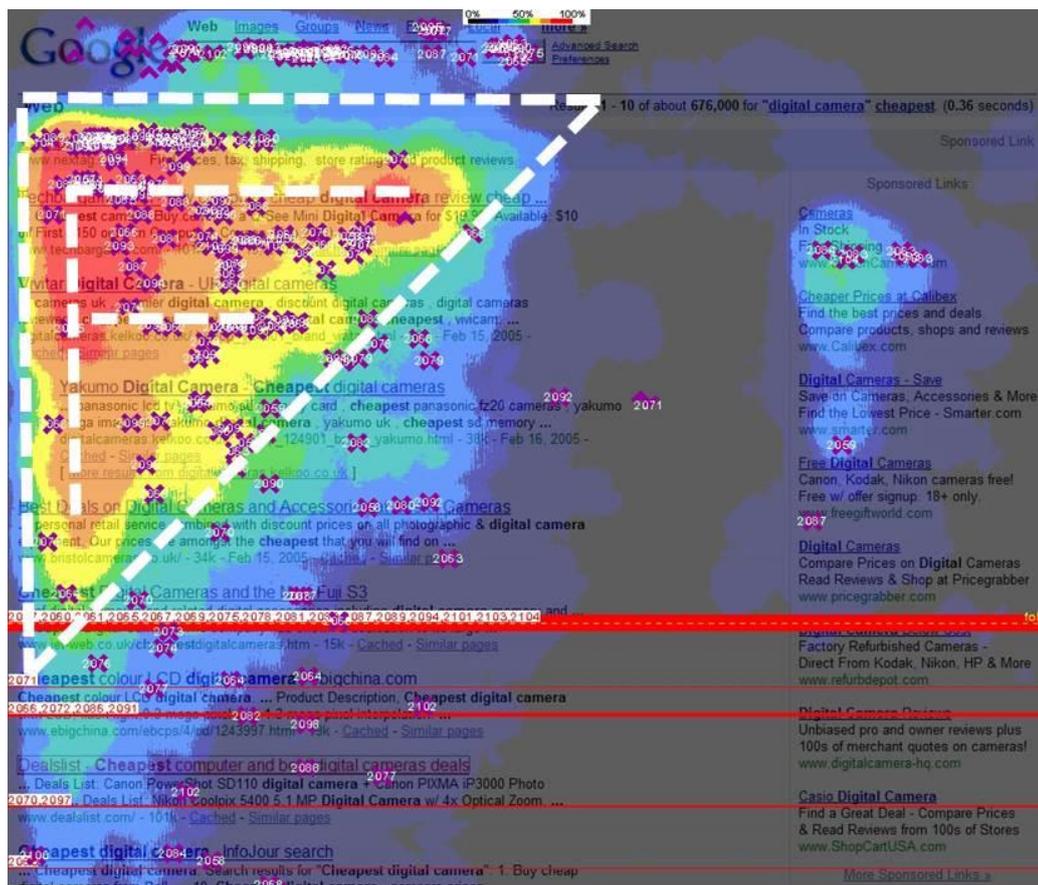


Abbildung 1: Das "Golden Triangle" und das "F-Pattern" in der Heatmap (vgl. ENQUIRO 2005)

In den nachfolgenden Enquiro-Studien von Hotchkiss et al. in den Jahren 2006 und 2007 wurde an die vorangegangene Studie angeknüpft, was das Untersuchungsdesign und die Forschungsfragen betrifft. Allerdings wurde die Evaluierung der SERP durch den Nutzer auf weitere Suchmaschinenanbieter wie Ask, MSN und Yahoo erweitert. Ziel war es herauszufinden, ob bei den Suchmaschinen Unterschiede beim Nutzerverhalten festzustellen sind und ob der „Golden Triangle“ Effekt und die F-förmige Betrachtungsweise der Suchergebnisse in den weiteren Erhebungen zu bestätigen sind.

Die Ergebnisse der Studien zeigten, dass durch die ab 2007 von Google und andere Suchmaschinen praktizierte Einbindung von Universal Search Ergebnissen (etwa aus der Bildersuche) und Shortcuts (direkte Fakteninformationen, z.B. Wetteranzeige) in die organischen Suchergebnisse von keiner listenartigen Ergebnispräsentation mehr gesprochen werden kann. Durch die veränderte Darstellung der SERP konnte auch ein vom ersten Studienergebnis abweichendes Evaluationsverhalten der Probanden festgestellt werden. Wurde etwa ein Element aus der Bildersuche auf einer SERP als Treffer angezeigt, orientierte sich der Blick der Testperson daran und folgte von dort aus nach oben, unten und nach rechts. Die Aufmerksamkeit der Probanden konzentrierte sich nicht mehr auf einen bestimmten Bereich, sondern war breiter auf der SERP verteilt (vgl. ENQUIRO 2006, 2008). Die Heatmaps aus den Untersuchungen visualisierten dieses Blickverhalten durch ein vom „Golden Triangle“ und „F-Pattern“ abweichendes Muster (vgl. Abbildung 2).

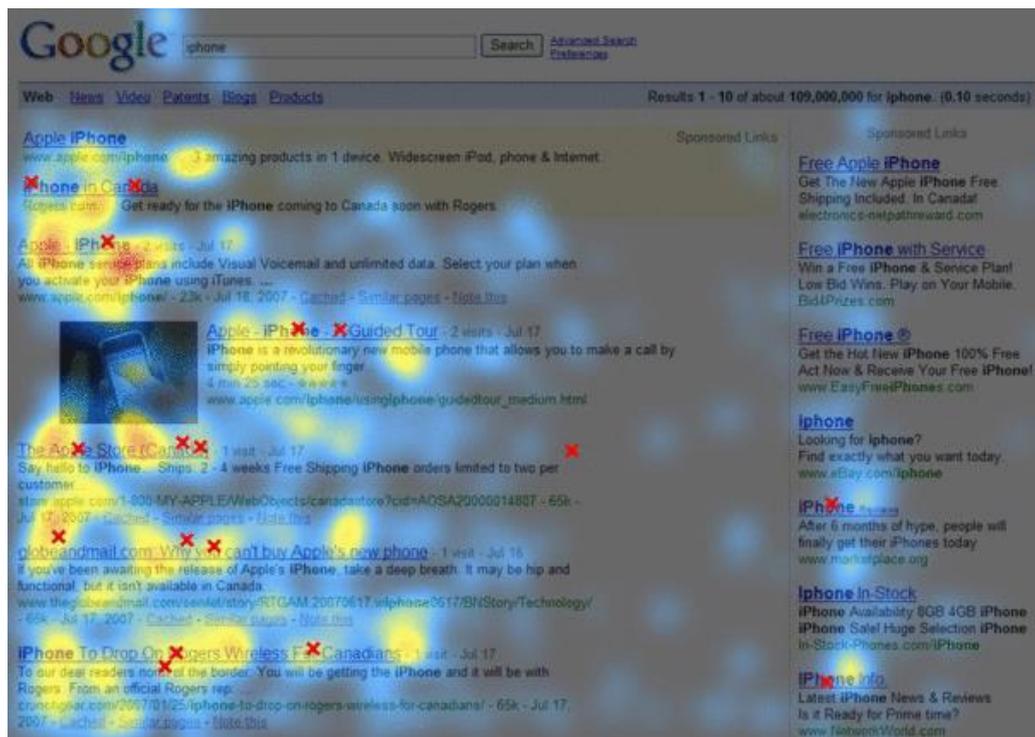


Abbildung 2: Durch die Einbindung von Suchergebnissen anderer Quellen entsteht eine diffusere Wahrnehmung der SERP durch die Nutzer (vgl. ENQUIRO 2008)

Das veränderte Evaluationsverhalten der Nutzer bei der Einbindung von Universal Search Ergebnissen und Shortcuts auf der SERP bestätigen u.a. die Untersuchungsergebnisse der Studien von usability.de und González-Caro und Marcos. Hervorgehoben wurde in ihnen, dass viele Universal Search Elemente selbst, wie etwa ein Maps-Bild, nicht besonders intensiv betrachtet werden, dafür aber die darunter liegenden Brancheneinträge. Treffer, die unmittelbar nach solchen Elementen gelistet sind, können so von einer erhöhten Aufmerksamkeit der Nutzer profitieren (vgl. USABILITY.DE 2009, S.7, GONZÁLEZ-CARO/MARCOS 2010, S.5).

Zusätzlich ergaben die Enquiro-Studien von 2006 und 2007, dass die Hervorhebung des als Suchbegriff verwendeten Keywords, vor allem im Titel des Suchergebnisses, sich positiv auf die Evaluierung der Treffer durch die Nutzer auswirkt. Eine solche optische Hervorhebung der benutzten Suchwörter als „Scent of Information“ hilft den Nutzern beim schnellen Scannen der ausgegebenen Treffer auf die erhoffte Information (vgl. ENQUIRO 2006, 2008).

Was die Unterschiede zwischen den einzelnen Suchmaschinen betrifft, konnte festgestellt werden, dass die Wahrnehmung der SERP bei allen untersuchten Suchmaschinen ähnlich ist. Das Evaluationsverhalten der Nutzer der Suchmaschine Google erfolgte allerdings schneller und weniger aufwendig, als bei den anderen Suchmaschinen (vgl. ENQUIRO 2006, 2008). Auch in einer 2007 von van Gisbergen, van der Most und Aelen durchgeführten Studie wurde festgestellt, dass Probanden bei der Suchmaschine Google sich im Vergleich anderen Suchmaschinenanbietern schneller zum Klicken eines Suchergebnis auf der SERP entschieden haben (vgl. GIESBERGEN/MOOST/AELEN 2007, S.5).

### **2.2.1.3 Die Studie von van Gisbergen et al.**

Die Eye Tracking Studie von Giesbergen et al. diente der differenzierteren Betrachtung der Suchanfragetypen nach Broder. Die zentrale Forschungsfrage war, ob es Unterschiede bei der Wahrnehmung der SERP zwischen informationsorientiert und transaktionsorientiert suchenden Nutzern gibt. Insgesamt nahmen 50 Probanden an der Studie teil denen jeweils 5 SERP der Suchmaschinen Google, MSN, Ilse, Lycos und Kobala gezeigt wurden. Die Hälfte der Testpersonen betrachteten SERP auf informationsorientierte Suchanfragen, der anderen Hälfte wurden SERP gezeigt, die auf transaktionsorientierte Suchanfragen ausgegeben worden sind (vgl. GIESBERGEN/MOOST/AELEN 2007, S.4).

Van Giesbergen et al. konnten feststellen, dass grundsätzlich transaktionsorientiert suchende Testpersonen sich mehr Zeit für die Betrachtung der SERP nahmen (11,4 sec) und insgesamt mehrere Suchergebnisse sahen (9,9), als es bei informationsorientiert suchenden Probanden der Fall war (9,4 sec und 8,5 Suchergebnisse) (vgl. GIESBERGEN/MOOST/AELEN 2007, S.5).

Interessant waren auch die Erkenntnisse aus einer Befragung der Testpersonen, die nach dem Eye Tracking Experiment erfolgte um die Hintergründe zur Klickentscheidung der Probanden zu ermitteln. Insgesamt

konnten 7 Faktoren bestimmt werden, die sich positiv auf die Klickentscheidung der Nutzer auswirken (vgl. GIESBERGEN/MOOST/AELEN 2007, S.6):

- 1) Die Vertrautheit und Zuverlässigkeit des Webdokuments beurteilt nach Keywords wie Markennamen
- 2) Das von der Suchmaschine vergebene Ranking des Webdokuments
- 3) Das Vorkommen und Hervorheben des Suchbegriffs im Suchergebnis
- 4) Das Vorkommen von positiven Nutzerbewertungen und – Kommentaren im Suchergebnis
- 5) Hinweise auf einen niedrigen Preis im Suchergebnis bei Suchanfragen mit einer Kaufabsicht
- 6) Hinweise auf eine große Produktvielfalt im Suchergebnis bei Suchanfragen mit einer Kaufabsicht
- 7) Das Verzichten von zu aggressiver Werbesprache im Suchergebnis

#### **2.2.1.4 Die Studie von González-Caro & Marcos**

Eine weitere Studie zur Analyse des Nutzerverhaltens auf den SERP in Hinblick auf die drei Suchanfragetypen, stellt die Eye Tracking Untersuchung von González-Caro und Marcos aus dem Jahr 2010 dar, wobei bei dieser Studie auch navigationsorientierte Suchanfragen berücksichtigt worden sind. An dem Experiment nahmen 58 Probanden teil, denen jeweils 6 Suchaufgaben zur Lösung mit einer Suchmaschine gegeben wurden. Insgesamt wurden 18 Suchanfragen zu allgemeinen Themen vorformuliert, 11 informationsorientierte, 3 navigationsorientierte und 4 transaktionsorientierte, die durch die Probanden zur Suche verwendet wurden. Die Hälfte der Testpersonen benutzte die Suchmaschine Google, die andere Hälfte Yahoo Search (GONZALEZ-CARO/MARCOS 2010, S.2).

Grundsätzlich konnte festgestellt werden, dass das Ranking der Treffer für das Evaluationsverhalten der Nutzer aller Suchintentionen eine wichtige Rolle spielt. Sowohl bei den informations-, transaktions- als auch bei den navigationsorientierten Suchanfragen wurden die Treffer der organischen

Suchergebnisse auf der ersten Position am intensivsten wahrgenommen. Die AdWords Anzeigen wurden bei informations- und navigationsorientierten Suchanfragen kaum beachtet, fanden aber bei transaktionsorientierten Suchanfragen, bei denen durchaus eine Kaufabsicht im Vordergrund der Suche stehen kann, mehr Beachtung durch die Testpersonen (GONZALEZ-CARO/MARCOS 2010, S.4).

Bei einer detaillierten Betrachtung der Ergebnisse, kam auch heraus, dass das Evaluationsverhalten der Nutzer auf der SERP durchaus von der Suchintention abhängig ist: Im Vergleich mit der Häufigkeit und der Dauer der Fixierungen einer SERP bei navigationsorientierten Suchanfragen sind diese bei informations- und transaktionsorientierten Suchanfragen länger und zahlreicher. Es konnte also festgestellt werden, dass bei der informations- und transaktionsorientierten Suche eine intensivere Betrachtung der SERP für die Evaluation der Suchergebnisse nötig ist (GONZALEZ-CARO/MARCOS 2010), S.5.

Die Ergebnisse von van Giesbergen et al., dass bei transaktionsorientierten Suchanfragen die Betrachtungsdauer der SERP länger ist als bei informationsorientierten, konnte allerdings nicht bestätigt werden. Eine dritte Studie zur Wahrnehmung der SERP durch die Nutzer in Hinblick auf die Suchabsicht nach Broder wurde von Ladstätter im Jahr 2009 durchgeführt. Diese hat zu den Erkenntnissen von van Giesbergen widersprüchliche Ergebnisse gezeigt (vgl. LADSTÄTTER 2009, S.2), sodass, was die Betrachtungsweise der SERP differenziert auf die drei Suchanfragetypen angeht, keine eindeutigen Verhaltensmuster festgestellt werden können.

González-Caro und Marcos konnten auch Unterschiede in der Wahrnehmung der einzelnen Suchergebniselemente, bestehend aus Suchergebnistitel, -Snippet und -URL, feststellen. Diese wurden als einzelne AOI der SERP definiert, die durch die Probanden je nach Suchintention, bei der Evaluierung der Suchergebnisse unterschiedlich stark einbezogen wurden. Grundsätzlich wurde die Suchergebnisbeschreibung am intensivsten wahrgenommen, was mit der Größe dieses Elements zu begründen ist.

Bei informationsorientierten Suchanfragen lag der Fokus auf der Betrachtung der Snippets, was daran liegt, dass im Beschreibungstext der Suchergebnisse häufig direkt nach den erwünschten Informationen gesucht wird. Der Titel und die URL der Suchergebnisse waren dagegen bei informations- und vor allem transaktionsorientierten Suchanfragen von größerer Bedeutung, als bei den informationsorientierten (GONZALEZ-CARO/MARCOS 2010, S.4).

## **2.2.2 Studien zu manipulierten SERP**

Um die Ergebnisse aus den bereits vorgestellten Studien zu überprüfen, wurden auch zahlreiche Eye Tracking Untersuchungen mit manipulierten SERP durchgeführt, bei denen etwa die Positionen der Treffer verändert oder einzelne Suchergebniselemente anders dargestellt wurden. Ziel war es herauszufinden, ob bei solchen SERP das Betrachtungs- und Selektionsverhalten der Nutzer ein anderes ist. Die Frage, inwiefern sich die Nutzer auf das Relevanzranking der Suchmaschine verlassen, stand dabei im Vordergrund dieser Studien.

### **2.2.2.1 Die Studie von Pan et al.**

Eine Eye Tracking Studie mit manipulierten Google-SERP wurde 2007 von Pan et al. durchgeführt. Zweck der Arbeit war es herauszufinden, ob die Fokussierung der Nutzer auf die Treffer der ersten beiden Positionen auf die tatsächliche Relevanzbewertung der Nutzer oder durch den Einfluss des Relevanzrankings der Suchmaschine zurückzuführen ist. Dazu wurde das Ranking auf der SERP manipuliert um zu überprüfen, ob dies Auswirkungen auf das Selektionsverhalten der Nutzer hat.

An der Studie nahmen 16 Probanden teil, denen jeweils 10 SERP ausgegeben auf 5 navigationsorientierte und 5 informationsorientierte Suchaufgaben gezeigt worden sind. Einige SERP wurden unverändert ausgegeben, bei einigen wurde das Ranking der Treffer zuvor umgekehrt (Position 1 auf 10, Position 2 auf 9 usw.) und bei einem dritten Teil der SERP wurde

das Ranking der Treffer getauscht (Position 1 mit 2, Position 3 mit 4 usw.) (vgl. PAN et al. 2007, S.802).

Es konnte festgestellt werden, dass beim unveränderten Ranking der SERP in 85% der Fälle die Suchaufgaben gelöst werden konnten, ohne dass Suchanfragen reformuliert werden mussten. Bei getauschten Positionen waren es noch 80%, während bei den SERP mit umgekehrten Positionen die Probanden größere Schwierigkeiten hatten, die Suchaufgaben zu lösen. 38% konnten bei den umgekehrt gerankten Treffern kein zufriedenstellendes Ergebnis erzielen. Auch hinsichtlich der Betrachtungsdauer gab es Unterschiede bei den manipulierten SERP. Bei den umgekehrt gerankten Treffern brauchten die Probanden 10,9 Sekunden zum Betrachten der Suchergebnisseite, während es in den anderen beiden Fällen jeweils unter 6,5 Sekunden waren. Auch wurde die SERP insgesamt häufiger fixiert und mehr Treffer gesichtet: während bei dem normal und getauscht angezeigten Ranking durchschnittlich unter 2,7 Treffer betrachtet wurden, waren es beim umgekehrten Ranking 3,8 Treffer (vgl. PAN et al. 2007, S.812). Allerdings hat der zusätzliche Zeitaufwand bei der Sichtung der SERP mit umgekehrtem Ranking zu keinen besseren Rechercheergebnissen geführt (vgl. PAN et al. 2007, S.812).

Insgesamt konnten im Normalfall der SERP die Ergebnisse von Granka et al. bestätigt werden, dass die ersten beiden Treffer die größte Aufmerksamkeit erhielten, wobei der erste am häufigsten geklickt worden ist. Es wird deutlich, dass auch die von Granka et al. beobachtete starke Abnahme der Sichtung und Selektion der Treffer ab Position 5 bestätigt worden ist (vgl. PAN et al. 2007, S.816).

Bei getauschten Positionen auf der SERP waren die Blicke der Probanden ähnlich stark auf die ersten beiden Treffer fixiert, wobei auch in diesem Fall die meisten Klicks der erste Treffer, im Original-Ranking also Position 2, erhielt. So konnte durchaus eine Beeinflussung durch das Ranking der Suchmaschine auf das Selektionsverhalten der Nutzer beobachtet wer-

den. Ansonsten konnten kaum Unterschiede im Nutzerverhalten zu der unmanipulierten SERP festgestellt werden (vgl. PAN et al. 2007, S.816).

Bei der umgekehrten Positionierung der Suchergebnisse konnten stärkere Abweichungen vom Nutzerverhalten im Normalfall beobachtet werden: die Treffer, die ursprünglich die Position 1 und 2 hatten, erhielten eindeutig weniger Klicks. Sowohl das Sichten als auch das Klicken der Treffer war bis zur Position 10 im Vergleich zum Normalfall gleichmäßiger verteilt (vgl. PAN et al. 2007 S.818).

Dieses Verhalten der Probanden weist darauf hin, dass sich die Nutzer bis zu einem gewissen Maß eher vom Ranking der Suchmaschine beeinflussen als dass sie sich von eigenen Relevanzbeurteilungen leiten lassen. Sie vertrauten dem Relevanzranking der Suchmaschine und klickten scheinbar gut gerankte Treffer, auch wenn diese zur Lösung der Suchaufgabe weniger relevant waren. Das Verhalten der Testpersonen beim umgekehrten Ranking zeigte allerdings, dass die Probanden trotzdem nicht blind das erste Suchergebnis geklickt haben. Allerdings führte diese Manipulation der SERP dazu, dass die meisten sich zwar mehr Zeit zum Betrachten der Ergebnisseite nahmen, allerdings dennoch die Suche nicht erfolgreich abschließen konnten. Es kann also durchaus davon gesprochen werden, dass die Nutzer auf das Relevanzranking der Suchmaschine vertrauen und die Position der Treffer einen großen Einfluss auf das Evaluationsverhalten der Nutzer hat (vgl. PAN et al. 2007, S.818).

Zu ähnlichen Ergebnissen führte eine 2008 von Keane, O'Brien und Smyth durchgeführte Studie, bei der 30 Teilnehmern 16 Fragen gestellt wurden, die sie mithilfe der Suchmaschine Google beantworten sollten. Auf die von den Probanden formulierten Suchanfragen wurden abwechselnd SERP mit umgekehrten Rankingpositionen und mit Originalranking angezeigt. Während das erste Suchergebnis im Normalfall von 70% der Probanden geklickt wurde, haben es im umgekehrten Ranking nur 10% aller Teilnehmer ausgewählt. Treffer der Positionen 9 und 10 wurden dagegen bei unmanipulierten SERP nur von 2 % geklickt, während es bei

den manipulierten SERP 13% und 41% waren (vgl. KEANE/O'BRIEN/SMYTH 2008, S.5).

### **2.2.2.2 Die Studie von Cutrell & Guan**

Welche Auswirkungen eine Veränderung des Suchergebnisrankings auf das Evaluationsverhalten hat und ob es dabei Unterschiede zwischen informations- und navigationsorientiert suchenden Nutzern gibt, wurde ebenso in der Eye Tracking Studie von Cutrell und Guan im Jahr 2007 erforscht. Den 18 akquirierten Testpersonen wurden 12 Suchaufgaben beider Anfragetypen zu lösen mit der Suchmaschine MSN Search gegeben. Auf den ausgegebenen SERP wurde die Position des ersten Treffers im Relevanzranking der Suchmaschine variiert, um feststellen zu können, bei welcher Positionierung dieser Treffer von den Testpersonen noch wahrgenommen und geklickt wurde (vgl. CUTRELL/GUAN 2007, S.407).

Generell konnte wie bei Pan et al. festgestellt werden, dass die Nutzer mehr Zeit mit der Evaluierung der Treffer verbrachten, wenn der ursprünglich erste Treffer auf niedrige Positionen gesetzt wurde. Dies hat ihnen allerdings nicht dabei geholfen, zufriedenstellende Suchergebnisse zu finden. Während beim unmanipulierten Ranking 78% der navigations- und 89% der informationsorientiert suchenden Probanden den erstplatzierten Treffer anklickten, waren es nur 11% sobald dieser auf die Position 8 verschoben wurde. Dieser Effekt fiel insgesamt bei den informationsorientierten Suchanfragen stärker aus. Was die Wahrnehmung des ursprünglich am besten gerankten Treffers betraf, nahm diese stark ab, je weiter seine Position nach hinten verschoben wurde. Besonders bei informationsorientiert suchenden Probanden konnte eine geringe Bereitschaft, die SERP auf die hinteren Positionen hin zu evaluieren, festgestellt werden (vgl. CUTRELL/GUAN 2007, S.412).

Obwohl der bestgerankte Treffer verschoben auf die Positionen 2-4 noch durch die meisten Probanden gesichtet worden ist, nahm die Bereitschaft auf diesen zu klicken deutlich ab. Besonders bei den informationsorientiert

suchenden Testpersonen konnte dies festgestellt werden: Obwohl 94% den Treffer auf Position 2 wahrnahmen, haben nur 33% ihn tatsächlich angeklickt. Bei der Position 4 waren es ebenfalls 89% die bereit waren den Treffer zu sichten, allerdings nur 17% die das Suchergebnis ausgewählt haben (vgl. CUTRELL/GUAN 2007, S.412).

Durch die Manipulation der SERP hinsichtlich der Position der Treffer konnte festgestellt werden, dass durchaus ein Zusammenhang zwischen dem Rank der Suchergebnisse und der Evaluierung dieser durch den Nutzer besteht. Vor allem bei informationsorientiert Suchenden sank die Bereitschaft das ursprünglich bestgerankte Suchergebnis auszuwählen, so dass viele Suchaufgaben nicht zufriedenstellend gelöst werden konnten. Es scheint also ein großes Vertrauen in das Ranking der Suchmaschine zu bestehen (vgl. CUTRELL/GUAN 2007, S.414).

Welche Auswirkungen eine manipulierte Länge der Beschreibungstexte eines Suchergebnisses auf das Sichtungs- und Selektionsverhalten der Nutzer hat, wurde ebenfalls in der Studie erforscht. Dabei wurden den Probanden drei Typen von SERP gezeigt: Eine, bei der die Snippet der Suchergebnisse kurz (eine Zeile), eine bei der sie mittellang (zwei bis drei Zeilen) und schließlich Suchergebnisseiten mit langen Snippets (sechs bis sieben Zeilen) der Suchergebnisse (vgl. CUTRELL/GUAN 2007, S.411).

Das Experiment zeigte, dass sich bei navigationsorientierten Suchanfragen die Betrachtungsdauer pro Suchaufgabe bei kurzen und mittellangen Snippets nicht von der Betrachtungsdauer einer regulären Snippetlänge unterschied. Bei langen Snippets stieg diese allerdings um 10 Sekunden pro Suchaufgabe an. Bei informationsorientierten Suchanfragen mit langen Snippets dagegen brauchten die Probanden im Durchschnitt 24 Sekunden weniger Zeit um eine Suchaufgabe zu lösen. Dies könnte daran liegen, dass längere Snippets als Vorschau für die im Webdokument enthaltenen Informationen dienen und sich so positiv auf die Evaluierung durch die Nutzer auswirkten (vgl. CUTRELL/GUAN 2007, S 416).

Das weitere Klickverhalten der Testpersonen wurde auch hinsichtlich der Snippetlänge für beide Anfragetypen analysiert. Es konnte festgestellt werden, dass bei informationsorientierten Suchanfragen die längeren Snippets dazu führten, dass das bestgerankte Suchergebnis auch bei manipulierter Position um 11% häufiger geklickt wurde, als es bei den anderen Snippetlängen der Fall war. Im Gegensatz dazu wurde bei navigationorientierten Suchanfragen das beste Ergebnis schon bei mittlerer Snippetlänge nur noch von 47% statt von 58% bei normaler Länge geklickt (vgl. CUTRELL/GUAN 2007, S.413).

Es konnte somit geschlussfolgert werden, dass bei informationsorientierten Suchanfragen längere Snippets zu einem effektiveren Suchverhalten der Nutzer führte während bei navigationorientierten Suchanfragen der gegenteilige Effekt erzielt wurde (vgl. CUTRELL/GUAN 2007, S. 411).

### **2.3 Erkenntnisse über das Selektions- & Navigationsverhalten auf SERP**

Sowohl aus den Eye Tracking Studien mit nicht manipulierten SERP als auch aus denen mit veränderten Rankingpositionen und Snippetlängen können Aussagen über das Selektions- und Navigationsverhalten der Nutzer auf SERP gemacht werden. Allerdings ist hervorzuheben, dass diese Aussagen aufgrund der meist kleinen Probandengruppen und individueller Blickverhalten nicht repräsentativ auf alle Suchmaschinennutzer verallgemeinert werden können (vgl. FAULDRATH 2011).

Grundsätzlich zeigten alle Studien, dass unabhängig vom Blickmuster, welches sich durch die Einblendung von Universal Search Ergebnissen veränderte (vgl. ENQUIRO 2007, USABILITY.DE 2009), und von der Suchabsicht die ersten beiden Treffer am intensivsten von den Nutzern wahrgenommen werden und insbesondere das erste Suchergebnis am häufigsten geklickt wird (vgl. GRANKA et al. 2004, GONZÁLEZ-CARO/MARCOS 2010). Bereits ab der fünften Position nahm die Wahrnehmung der Treffer deutlich ab, was damit erklärt wird, dass die Nutzer

selten bereit sind auf der SERP nach unten zu scrollen. Dies kann damit begründet werden, dass die Nutzer dem Relevanzranking der Suchmaschine vertrauen. Die Studien mit manipulierten SERP bestätigten, dass die Position der Treffer eine entscheidende Rolle bei der Evaluierung der Suchergebnisse durch die Nutzer spielen und diese bei der Suche weniger erfolgreich sind, wenn die Position des bestgerankten Treffers verändert wurde (vgl. PAN et al 2007, KEANE/O'BRIEN/SMYTH 2008, CUTRELL/GUAN 2007). Somit kann das Selektions- und Navigationsverhalten der Nutzer auf der SERP als ein unbewusst durch das Ranking der Treffer beeinflusster Vorgang zur Befriedigung des Informationsbedürfnisses angesehen werden.

In Hinblick auf diese Arbeit bestätigen die angeführten Eye Tracking Studien, dass niedrig gerankte Suchergebnisse in den meisten Fällen weder durch die Nutzer gesichtet, evaluiert noch geklickt werden. Eine Relevanzanalyse der Webdokumente hinterer Trefferpositionen durch einen Retrievaltest würde zu Erkenntnissen über das Relevanzempfinden der Nutzer führen, ohne dass diese durch die Rankingpositionen der vorgelegten Webdokumente beeinflusst sind. So können die Relevanzbewertungen der Nutzer sowohl von vorderen als auch von hinteren Positionen verglichen werden. Vor dem Hintergrund der Suchmaschinenqualität kann so die Retrievaleffektivität überprüft und das systemeigene Ranking kontrolliert werden während für das Forschungsfeld des Nutzerverhaltens die Evaluierung der Nutzer gerechtfertigt oder in Frage gestellt werden kann.

### 3 Retrievaltests zur Suchmaschinenevaluierung

Die vorgestellten Eye Tracking Studien zeigen, dass das Selektions- und Navigationsverhalten der Suchmaschinennutzer auf der SERP von dem Relevanzranking der Suchmaschinen beeinflusst werden und durchaus Bedarf besteht, die Relevanz niedrig geranker Webdokumente als Qualitätsnachweis der Suchmaschine Google mit einem Retrievaltest zu ermitteln. Im Zentrum dieser Evaluationsmethode steht die Relevanzbewertung der Webdokumente durch Juroren. Hervorzuheben ist, dass diese nicht als objektive Einschätzung bezeichnet und der Begriff Relevanz nicht eindeutig definiert werden kann (vgl. QUIRMBACH 2012, S.132). Relevanz stellt somit sowohl den Ausgangspunkt zur Messung der Qualität durch Retrievaltests, als auch den Hauptkritikpunkt dieser Methode dar. Deshalb soll im nachfolgenden Kapitel die Diskussion in der Fachliteratur um den Begriff der Relevanz erläutert werden, bevor der Retrievaltest mit den in der Suchmaschinenforschung etablierten Methoden und Standards vorgestellt, einige Studien erläutert und die Ergebnisse daraus in Hinblick auf die eigene Studie zusammengefasst werden.

#### 3.1 Der Begriff der Relevanz

Relevanz wird in Retrievaltests als objektives Bewertungsmaß eingesetzt und stellt somit einen wichtigen Bestandteil der Qualitätsmessung in Suchmaschinen dar, obwohl die Definition des Begriffs nicht eindeutig festgelegt und Relevanzbewertungen selbst durch die Juroren subjektiv getroffen werden. Einer der führenden Experten in der Diskussion um den Begriff und die Messung von Relevanz drückt die bilaterale Bedeutung für die Informationswissenschaft folgendermaßen aus: „relevance became a key notion (and key headache) in information science“ (SARACEVIC 1999, S.1058).

Wolfgang G. Stock versucht Relevanz als Relation zwischen dem objektiven Informationsbedarf des Nutzers und dem gefundenen Dokument, bzw.

des darin enthaltendem Wissen, zu bezeichnen. Ein Dokument ist demnach relevant, wenn es dem Nutzer:

- bei der Vorbereitung einer Entscheidung hilft
- eine Wissenslücke schließt
- und als Frühwarnfunktion dient (vgl. STOCK 2007, S.69)

Was die Bewertung von Relevanz betrifft, kann diese nach dem binären Ansatz erfolgen, bei dem ein Dokument entweder als relevant oder nicht relevant bewertet, oder nach einer differenzierteren Skalenbewertung, bei der Werte zwischen relevant und nicht relevant angenommen werden können. Untersuchungen haben gezeigt, dass Laiennutzer eher dazu tendieren die Relevanz von Dokumenten in Intervallen zu bewerten, während Information Professionals eher nach dem binären Ansatz urteilen können (vgl. STOCK 2007, S.71). Die Relevanzbewertungen an sich bleiben allerdings von den subjektiven Eigenschaften der Urteilenden, wie deren Wissensstand und kognitive Fähigkeiten abhängig. Wann ein Webdokument als relevant bewertet wird, ist somit auf Entscheidungen nach individuellen Kriterien der Juroren zurückzuführen (vgl. GRIESBAUM 2000, S.67).

Durch die Problematik des Begriffs Relevanz müssen Retrievaltests umso sorgfältiger geplant um umgesetzt werden, etwa um den Juroren das Prinzip der Relevanzbewertung gleichermaßen zu erklären. Im Folgenden sollen die Standards für einen validen Retrievaltest vorgestellt werden, um trotz des uneinheitlich definierten Relevanzbegriff und die subjektiv und individuell durch die Juroren empfundenen Relevanzbewertungen möglichst aussagekräftige Ergebnisse zu erhalten.

### **3.2 Methoden und Standards für Retrievaltests**

Retrievaltests messen die Effektivität von Information-Retrieval-Systemen auf Suchanfragen der Nutzer relevante Dokumente auszugeben und sind ein wichtiges Mittel zur Messung ihrer Qualität. Was den Aufbau und die Vorgehensweise betrifft, haben sich Standards herausgebildet die im Folgenden erläutert werden.

### 3.2.1 Das Cranfield Modell

Das bereits 1960 von Cleverdon am Cranfield College etablierte Cranfield Modell stellt ein führendes Paradigma bei der Evaluation von Information-Retrieval-Systemen dar und kann auch heute als Basis für die Konzeption von Retrieval Tests angesehen werden. Nach diesem Modell müssen die Testkollektionen grundsätzlich als einem Dokumentkorpus, Topics (einem Katalog von Informationsbedürfnissen bzw. Suchanfragen) und den dazugehörigen Relevanzurteilen der Juroren bestehen. Werden die Dokumente auf die Topics hin ausgegeben, sollen diese von unabhängigen Juroren auf ihre Relevanz hin bewertet werden. Dabei soll jedes Dokument nur einem Juror vorgelegt werden. Daraufhin kann die Retrievaleffektivität des Systems durch die Maße Recall und Precision analysiert werden (vgl. CLEVERDON/MILLS/KEEN 1966, S.9).

### 3.2.2 Testaufbau nach TREC

Die meisten Retrievaltests folgen dem TREC-Testaufbau, der sich an dem Cranfield Modell orientiert. Dabei werden Suchanfragen an eine oder mehrere Suchmaschinen geschickt, die zurückgegebenen Webdokumente bis zu einer bestimmten Trefferposition (häufig bis Position 10) anonymisiert, gemischt und anschließend durch die akquirierten Juroren auf ihre Relevanz hin bewertet (vgl. VOORHEES/HARMAN 2005, S.563). Die Beschränkung auf eine gewisse Anzahl von Webdokumenten ist unvermeidbar, da die Bewertung aller durch die Suchmaschine gefundenen Treffer aufgrund der meist großen Treffermengen nicht umsetzbar ist (vgl. LEWANDOWSKI 2011, S.206). Die Webdokumente werden nach der Bewertung ihren Rankingpositionen und Suchmaschinen zugeordnet, um mit den Relevanzurteilen der Juroren die Retrievaleffektivität beurteilen oder Vergleiche zwischen den verschiedenen Suchmaschinen ziehen zu können. Häufig werden dabei Kennzahlen wie Precision und Recall herangezogen (vgl. VOORHEES/HARMAN 2005, S.565).

Bei der TREC-Methodik wird von einem typischen Nutzer als „dedicated Searcher“ ausgegangen, der sich als Information Professional eine große Anzahl an ausgegebenen Dokumenten auf seine Suchanfrage wünscht und auch bereit ist diese zu sichten (vgl. VOORHEES/HARMAN 2005, S.563). Erkenntnisse aus der Nutzerforschung, die vom Laiennutzer als Standardfall ausgehen, der weder viel Aufwand und Zeit bei der Formulierung seiner Suchanfrage noch beim Evaluieren der Treffer auf der SERP investiert (vgl. Kapitel 2.2), zeigen allerdings ein gegenteiliges Bild vom typischen Suchmaschinennutzer. Die Tatsache, dass Retrievaltests nach dem TREC-Standard unabhängig vom realen Nutzerverhalten konzipiert werden, kann durchaus als ein Kritikpunkt angesehen werden.

Ebenso problematisch ist beim Vergleich mehrerer Suchmaschinen das Erstellen von Testkollektionen, die gebildet werden um zum einen für den Test Suchanfragen aus demselben Index zu erhalten und zum anderen, da ein direkter Zugang zum Index der jeweiligen Suchmaschinen nicht gegeben ist. In der Praxis kann allerdings davon ausgegangen werden, dass die verschiedenen Suchmaschinen sich in ihrer Indexgröße und -qualität unterscheiden (vgl. LEWANDOWSKI 2011, S.204).

Nichtsdestotrotz stellt der TREC-Standard seit 1992 eine erprobte Methode für die großflächige Evaluation von Information-Retrieval-Systemen dar und wird bis heute verwendet.

### **3.2.3 Konzeption eines Retrievaltests nach Tague-Sutcliffe**

Neben dem Aufbau und der Durchführung des Retrievaltests sind zuvor grundlegende Fragen zur Konzeption dessen zu klären. Tague-Sutcliffe hat 10 Fragen zusammengestellt, die für eine begründete methodische Konzeption des Tests beantwortet werden sollten:

1. Testen oder nicht testen? Ein Test sollte demnach nur dann durchgeführt werden, wenn durch die Ergebnisse ein Erkenntnisfortschritt zu erwarten ist.

2. Wie Testen? Sin neue Erkenntnisse durch den Test zu erwarten, sollte die Methode des Tests festgelegt werden.
3. Welche Variablen sollen bestimmt werden? Die auf ihre Ausprägung zu testenden Merkmalsträger sollten klar definiert werden.
4. Welche Datenbank soll benutzt werden? Das Information-Retrieval-System, in diesem Fall die Suchmaschine, muss ausgewählt werden.
5. Welche Suchanfragen sollen gestellt werden? Um im Test reale Informationsbedürfnisse darzustellen, sollten auch reale Suchanfragen verwendet werden.
6. Wie sollen die Suchanfragen durchgeführt werden? Bevor die Webdokumente den Juroren vorgelegt werden können, müssen diese erst gefunden und gespeichert werden.
7. Wie erfolgt die Testanordnung? Es muss festgelegt werden, wie die Juroren die Webdokumente sichten und bewerten sollen.
8. Wie werden die Daten erhoben? Die Wahl sollte getroffen werden, wie die Relevanzurteile der Juroren festzuhalten sind.
9. Wie werden die Daten ausgewertet? Die statistische Auswertung der Daten sollte ebenfalls durchdacht sein.
10. Wie werden die Ergebnisse präsentiert? Die Form, in welcher die Ergebnisse des Tests dargestellt werden sollen, sollte geklärt werden.

Obwohl diese Fragen nicht direkt auf einen Retrievaltest für Suchmaschinen bezogen worden sind, stellen sie einen wichtigen Schritt bei der Planung und Konzeption dessen dar (vgl. TAGUE-SUTCLIFFE 1992, S.473). Sind diese beantwortet, kann die Testdurchführung vorbereitet werden.

### **3.2.4 Kriterien für Retrievaltests bei Suchmaschinen nach Hawking et al.**

Weitere 5 Kriterien für einen validen Retrievaltest sind in den Ausführungen von Hawking et al. zu finden. Sie entstanden in Anlehnung an die Arbeit von Gordon und Pathak (vgl. GORDON/PATHAK 1999, S.141) und

berücksichtigen die Eigenschaften der Suchmaschinen, die diese von anderen Information-Retrieval-Systemen unterscheiden. (vgl. HAWKING et al. 2001, S. 33 ff.).

1. Reale Informationsbedürfnisse von Nutzern sollen abgebildet werden. Dies ist wichtig, um die Ergebnisse des Tests durch künstlich generierte Suchanfragen nicht zu verfälschen.
2. Das Informationsbedürfnis soll den Juroren mitgeteilt werden. Vor allem, wenn die gesichteten Webdokumente nicht aus der eigenen Motivation der Juroren gesucht wurden, sollten sie sich durch Hintergrundinformationen zur Suche in die Situation hineinversetzen können.
3. Es soll eine große Anzahl von Suchmaschinen genutzt werden. Dies bezieht sich auf Tests, bei denen der Vergleich von Suchmaschinen im Vordergrund steht.
4. Es sollen wichtige Suchmaschinen involviert werden. Die Wahl der zu testenden Suchmaschinen sollte an der Popularität und den Marktanteilen dieser getroffen werden.
5. Die Untersuchung soll gut und sorgfältig aufgebaut und durchgeführt werden.

### **3.2.5 Framework für Retrievaltests nach Lewandowski**

Lewandowski bietet einen kompletten Überblick für die Evaluierung von Suchmaschinen durch moderne Retrievaltests in Form eines fünfteiligen Frameworks (vgl. LEWANDOWSKI 2012, S.8).

Im ersten Teil stehen, nach der Wahl der zu testenden Suchmaschine, die Suchanfragen für den Test im Vordergrund. Dabei soll pro Test der Fokus auf jeweils einem Anfragetyp liegen und inhaltlich vielfältige und das reale Informationsbedürfnis widerspiegelnde Themen gewählt werden, die etwa aus Logfiles entnommen werden können. Je mehr Suchanfragen beim Test verwendet werden, desto besser. Allerdings ist die begrenzte Anzahl der Juroren zu bedenken, sodass in der Praxis mit etwa 50 Suchanfragen

gearbeitet wird. Auch wird herausgestellt, dass die Formulierung einer Suchanfrage im Idealfall vom selben Probanden vorgenommen werden sollte, der anschließend das darauf ausgegebene Dokument bewertet, weil nur so das Informationsbedürfnis durch den Juror nachvollzogen werden kann. Da dies allerdings in den meisten Retrievaltests nicht praktiziert wird, sollte eine Suchanfragebeschreibung dem Juroren, die neben der Suchanfrage an sich auch Hintergrundinformationen zum Suchmotiv geben sollte. Auch ist zu berücksichtigen, dass Suchanfragen durchaus mehrdeutige Bedeutungen haben können, wie etwa die Suchanfrage „James Bond“. Diese kann sich beispielsweise auf Filme, Bücher, Biografien oder Schauspieler beziehen. Beim Retrievaltests sind also Informationen zur eindeutigen Bedeutung der Suchanfrage zu geben (vgl. LEWANDOWSKI 2012, S.8).

Im zweiten Teil wird die Zusammenstellung der Suchergebnisse und der zusätzlich durch Suchmaschinen gegebenen Informationen wie Trefferbeschreibungen und Positionen thematisiert. Was die Trefferbeschreibungen angeht, sollten diese bei der Evaluierung der Suchmaschine miteinbezogen werden, da sie meist eine große Rolle beim Selektionsverhalten der Nutzer auf der SERP haben. Das Suchergebnis als Webdokument und seine Position im Suchmaschinenranking ist zu erfassen, beim Vergleich mehrerer Suchmaschinen untereinander auch die Suchmaschine als Quelle des Webdokuments. Sowohl die Position als auch der Name der Suchmaschine sollte allerdings bei der Bewertung anonymisiert werden um Lerneffekte und Voreingenommenheit der Juroren zu vermeiden (vgl. LEWANDOWSKI 2012, S.8).

Im dritten Teil wird die Darstellung der Suchergebnisse diskutiert. Durch die von der Liste abweichende Darstellung der Treffer auf der SERP, etwa durch Shortcuts und Universal Search Ergebnisse, sollte der Typ des Suchergebnisses ebenfalls berücksichtigt werden. Auch ist zu beachten, dass Webdokumente ebenfalls differenziert werden können, etwa was ihre Herkunft von Blogs, Newsseiten und anderen Webseiten betrifft. Es sollte

also darauf geachtet werden, möglichst alle Typen im Test zu erfassen (vgl. LEWANDOWSKI 2012, S.8).

Wie die Webdokumente bzw. ihre Relevanz zu bewerten sind, wird im vierten Teil des Frameworks erläutert. Dabei spielt die Wahl der Juroren bereits eine wichtige Rolle. Diese sollten grundsätzlich die heterogene Zielgruppe der Suchmaschine von Laien bis Information Professionals widerspiegeln. Was die Anzahl der zu bewertenden Suchanfragen betrifft, sollte jeder Juror eine Suchanfrage erhalten und darauf mehrere Webdokumente bis zu einem bestimmten Cut Off Wert bewerten. Soll dagegen der Vergleich mehrerer Nutzergruppen durchgeführt werden, sollte eine Suchanfrage von Juroren aus jeder Gruppe bearbeitet werden. Der Cut Off Wert begrenzt die zur Bewertung vorgelegten Webdokumente auf eine bestimmte Position hin, die je nach Forschungsfrage und Ziel variieren sollte, häufig allerdings an das Nutzerverhalten bis zur 10 Position angelehnt wird. Was die Relevanzurteile betrifft, können diese binär oder in Skalen erfolgen, wobei eine differenzierte Analyse hinsichtlich hoch relevanter Webdokumente durch Skalenbewertungen zu erreichen ist (vgl. LEWANDOWSKI 2012, S.8).

Im letzten Teil des Frameworks wird die Datenanalyse thematisiert. Im Vordergrund der Retrievaltests steht die Ermittlung der Relevanz, die bezogen auf die Webdokumente durch die klassischen Information Retrieval Werte Recall und Precision gemessen werden können, wobei auch Fallout und Generality und weitere Kennzahlen herangezogen werden können. Werden Trefferbeschreibungen in den Test einbezogen, ist die Analyse derer in Zusammenhang mit den Webdokumenten ebenfalls durchzuführen. Wird etwa sowohl die Trefferbeschreibung an sich als auch das Webdokument für relevant bewertet, kann von einer idealen Retrieval-effektivität gesprochen werden. Wird dagegen die Trefferbeschreibung für relevant, das darauf folgende Webdokument allerdings für nicht relevant bewertet, ergibt sich eine andere Bewertung des Systems. Auch die Viel-

fältigkeit der Quellen der Suchergebnisse kann für die Bewertung der Suchmaschine herangezogen werden.

Hervorzuheben ist, dass die zu treffenden Entscheidungen Einfluss auf die Validität des Retrievaltests haben und je nach Forschungsfrage und Ziel individuell getroffen werden müssen. Eine veranschaulichte Zusammenfassung des ausführlichen Modells für die Durchführung und Aufbau von Retrievaltests zur Evaluierung von Suchmaschinen bietet die nachfolgende Abbildung (vgl. Abbildung 3).

### Framework for Evaluating the Retrieval Effectiveness of Search Engines

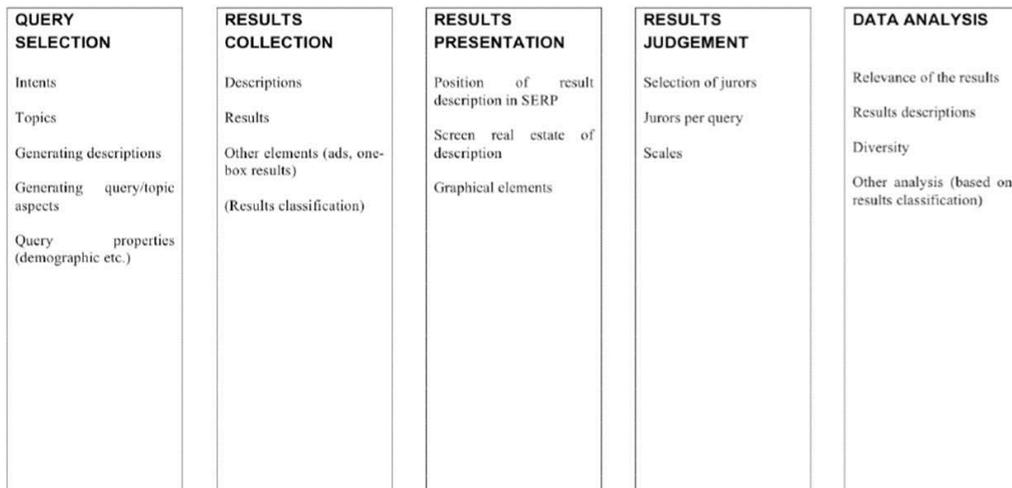


Abbildung 3: Zusammenfassung der 5 Schritte zum Retrievaltest (vgl. LEWANDOWSKI 2012)

#### 3.2.6 Maße und Kennzahlen für die Datenauswertung in Retrieval Tests

In den dargestellten Standards wurde die Wahl der Maße und Kennzahlen für die Datenanalyse in Retrievaltests thematisiert, die für die Evaluierung von Information Retrieval Systemen von Bedeutung sind. Es wurden unterschiedliche Kennzahlen entwickelt um die Qualität der Suchergebnistreffer zu bewerten. Diese wirken sich auf die Ergebnisse der Retrievaltests aus. Die gängigsten von ihnen im nachfolgenden Kapitel separat betrachtet und erläutert werden.

### 3.2.6.1 Die Maße Recall und Precision als Standard in Retrievaltests

Das am häufigsten in Retrievaltests jeglicher Information Retrieval Systeme verwendete und somit bedeutsamste Maß heißt Precision. Diese eignet sich zur Auswertung der Daten bei binären Bewertungsurteilen und gibt den Anteil der relevanten ausgegebenen Treffer aus der Gesamtheit aller ausgegebenen Treffer an. Berechnet wird sie indem aus allen ausgegebenen, auf ihre Relevanz hin beurteilten Treffer, die Anzahl der als relevant gewerteten, ausgezählt wird. Bei großen Treffermengen, wie es in Suchmaschinen meist der Fall ist, wird die Precision bis zum Cut Off Wert gemessen. Mathematisch lässt sich Precision folgendermaßen ausdrücken:

$$p = \frac{a}{a + b}$$

wobei  $p$  Precision,  $a$  Zahl der gefundenen relevanten Dokumente und  $b$  Zahl der gefundenen nicht relevanten Dokumente ist (vgl. LEWANDOWSKI 2012, S.213).

Ebenfalls als Standardmaß kann die Kennzahl Recall bezeichnet werden. Dieser wird in Bezug auf Suchmaschinen als der Anteil der relevanten ausgegebenen Treffer aus der Gesamtheit der insgesamt im Web vorhandenen relevanten Treffer bezeichnet. Hervorzuheben ist, dass sich der Recall so nur schwer messen lässt, da die Anzahl der im Web vorhandenen relevanten Webdokumente nicht exakt bestimmt werden kann. Die Formel für Recall sieht folgendermaßen aus:

$$r = \frac{a}{a + c}$$

wobei  $r$  Recall,  $a$  Zahl der gefundenen relevanten Dokumente und  $c$  Zahl der nicht gefundenen relevanten Dokumente ist (vgl. LEWANDOWSKI 2012, S.213).

Die Werte der beiden Kennzahlen können zwischen 0 und 1 liegen, wobei 1 als der beste und 0 als der schlechteste Wert betrachtet werden. Liegt die Precision etwa bei 1 bedeutet dies, dass alle gefundenen Webdokumente relevant sind, während bei 0 keins der gefundenen Webdokumente so einzustufen ist. Liegt der Recall bei 1, würde dies bedeuten, dass alle relevanten Webdokumente des Webs durch die Suchmaschine gefunden worden sind, während bei 0 kein relevantes Dokument gefunden wurde. Eine qualitativ hochwertige Suchmaschine sollte demnach sowohl einen hohen Recall als auch eine hohe Precision aufweisen.

Bei den Maßen Recall und Precision ist zu beachten, dass ihre Verwendung durchaus vom Anfragetyp der gebrauchten Suchanfragen abhängig gemacht werden sollte. Beim informationsorientierten Suchanfragen, bei denen eine bestimmte Menge von relevanten Ergebnissen durchaus gewünscht wird, können diese Kennzahlen zu aussagekräftigen Ergebnissen führen, während etwa bei navigationsorientierten Suchanfragen, bei denen nur ein bestimmtes Webdokument gefunden werden und als relevant einzustufen ist, andere Kennzahlen im Vordergrund stehen sollten (vgl. LEWANDOWSKI 2012, S.214).

Auch gibt es für beide Kennzahlen zwei Ansetze: die Makro- und Mikromittelwertbildung. Bei der Makrobewertung wird der Mittelwert über alle Recall und Precision Werte gebildet, also für alle Suchvorgänge. Die einzelnen Suchanfragen werden als Einheit betrachtet, sodass zuerst alle Precisionwerte pro Suchanfrage berechnet und dann die Werte der Suchanfragen gemittelt werden, damit jede Suchanfrage gleichwertig in die Bewertung eingezogen wird. Deshalb spricht man von einer nutzerorientierten Mittelwertbildung, die zeigt wie effektiv die Suchmaschine einzelne Suchanfragen beantworten kann. Bei der Mikrobewertung werden die Dokumente als Einheit angesehen. Unabhängig von der Trefferzahl der einzelnen Suchanfragen, wird das Verhältnis der relevanten Treffer zu allen Treffern berechnet. Dadurch wird jedes Dokument gleich der Bewertung hinzugezogen. Es wird eine Summe aus der Zahl der relevanten und nicht

relevanten Webdokumente aller Suchvorgänge gebildet und danach ein Recall und Precision Wert errechnet, also eine systemorientierte Mittelwertbildung. Diese zeigt die Effektivität der Suchmaschine bis auf einen Cut-off Wert über alle Suchanfragen hinweg.

Da Suchmaschinen mit gerankten Ergebnislisten arbeiten, wurden die Kennzahlen weiterentwickelt, um diesen Faktor bei der Messung in Retrievaltests zu berücksichtigen, etwa durch die Kennzahl Cut-off-Precision,  $p(k)$ . Hierbei wird nur ein Teil der vom System ausgegebenen Ergebnisse für die Berechnung verwendet, häufig die ersten 10 oder 20 Treffer, passend zu dem Verhalten der Nutzer, die in der Regel nur die ersten Suchergebnisse sichtet.

### 3.2.6.2 Weitere Kennzahlen

Weitere etablierte Kennzahlen bei Retrievaltests sind Fallout als der Anteil der ausgegebenen, aber nicht relevanten Treffer aus der Gesamtzahl aller nicht relevanten Treffer und Generality als der Anteil der relevanten Dokumente im Datenbestand. Von den zahlreichen weiteren Kennzahlen wie Median Measure, Excepted Search Length, Success und Satisfaction/Frustration wurden die meisten in der Praxis allerdings nur selten verwendet, sodass sich speziell für den Bereich der Suchmaschinenforschung kein Maß durchsetzen konnte. Es wird besonders deutlich, dass Kennzahlen entwickelt werden müssen, die einerseits Aussagen über einzelne Treffer als auch hinsichtlich der Trefferliste möglich machen (vgl. LEWANDOWSKI 2012, S.214-215).

Die vorgestellten Standards, Methoden und Kennzahlen für den Aufbau und die Datenauswertung in Retrievaltests zeigen, dass eine solche Evaluation von Suchmaschinen mit Aufwand verbunden ist. Um valide Ergebnisse zu erhalten, sind bei der Konzeption eines Retrievaltests wichtige Entscheidungen zu treffen, die je nach Forschungsfrage und Ziel individuell zu betrachten sind.

### 3.3 Stand der Forschung über Retrievaltests

In den letzten Jahren wurden zahlreiche Retrievaltests zur Evaluierung von Suchmaschinen durchgeführt. Einige dienen dem Vergleich der Retrievaleffektivität von Suchmaschinen mit anderen Information-Retrieval Systemen (vgl. BROPHY/BAWDEN 2005), andere wiederum wurden durchgeführt um die Qualität der populärsten Suchmaschinen untereinander zu vergleichen (vgl. GRIESBAUM/RITTBERGER/BEKAVAC 2002; VÉRONIS 2006; LEWANDOWSKI 2008B, 2011B). Letztere sollten häufig eine Begründung für die dominierende Rolle von Google am Suchmaschinenmarkt finden. Im Folgenden sollen einige auf informationsorientierte Suchanfragen bezogene Retrievaltests und ihre Ergebnisse vorgestellt werden, um zum einen das Verfahren des Retrievaltests an praktischen Beispielen zu verdeutlichen und um den aktuellen Forschungsstand auf dem Gebiet der Suchmaschinenforschung hinsichtlich der Suchmaschinenqualität aufzuzeigen. Welche Methoden und Kennzahlen verwendet und welchen Wert die Erkenntnisse in Hinblick auf die eigene Arbeit haben, soll beurteilt werden.

#### 3.3.1 Die Retrievaltests von Griesbaum

Ob es Unterschiede in der Qualität der Suchmaschinen von verschiedenen Diensteanbietern gibt, sollte unter anderem der Retrievaltest von Griesbaum aus dem Jahr 2002 beantworten. Dazu wurden die 4 derzeitig populärsten deutschsprachigen Suchmaschinen AltaVista.de, Fireball.de, Google.de und lycos.de ausgewählt. Getestet wurde größtenteils mit 56 Suchanfragen eingeschränkt auf die Themen Wissenschaft und Politik, sodass das typische Verhalten der Nutzer, zu weitaus vielfältigeren Themen zu recherchieren, nicht abgebildet wurde. Diese stammen aus TREC Testkollektionen und aus Auswertungen von realen Anfragen im Web. Die gefundenen Treffer wurden den 28 Juroren, von denen die meisten Studenten waren, anonymisiert zur Relevanzbewertung vorgelegt, sodass diese nicht wussten, welche Treffer von welcher Suchmaschine stammen

um unvoreingenommen zu bewerten. Jeder Juror bekam 2 Suchanfragen zum Beurteilen. 3 Relevanzbewertungen waren möglich: ein Webdokument konnte als relevant, nicht relevant oder als auf ein relevantes Dokument verweisend, bewertet werden. Letztere wurden in der Auswertung zu den relevanten Treffern gezählt, da sie das Informationsbedürfnis ebenfalls befriedigten konnten, wodurch ein binäres Werturteil entstehen konnte, allerdings insgesamt höhere Relevanzwerte zu erwarten waren, als es in den klassischen binären Bewertungssystem der Fall wäre (vgl. GRIESBAUM /RITTBERGER /BEKAVAC 2002, S.6).

Gemessen wurden die Precision der ersten 20 Trefferpositionen, also die Top 20 Precision, sowohl mit der Mikro als auch der Makromethode, und die Anzahl der insgesamt mit mindestens einem relevanten Treffer beantworteten Suchanfragen. Zur Absicherung der Validität der Ergebnisse wurden die Werte auf Signifikanz überprüft (vgl. GRIESBAUM /RITTBERGER /BEKAVAC 2002, S.7).

Insgesamt wurde so festgestellt, dass Google bei den meisten Werten am besten abgeschnitten hat. Von den maximal 1120 ausgegebenen Treffern der Suchmaschine wurden über 50% als relevant beurteilt, gefolgt von Lycos, während bei AltaVista und Fireball dieser Wert bei etwa einem Drittel lag (vgl. GRIESBAUM /RITTBERGER /BEKAVAC 2002, S.13).

Mit der Mean Average Precision (Mikromethode), veranschaulicht in Abbildung 13, konnte festgestellt werden, dass Google über alle Suchanfragen hinweg, auf Position 1 58,93% der Treffer für relevant bewertet wurden und bis einschließlich Position 20 kumuliert (Position 1-20), es 51% waren. Dieser Wert ist als auch Gesamt- also Top 20 Precision zu bezeichnen. Bei Lycos wurden 57% der Treffer auf Position 1 für relevant bewertet, während die Top 20 Precision bei 42% lag. Errechnet aus dem Durchschnitt der pro Rangplatz erreichten Precisionwerte der Suchmaschinen zur Mean Average Precision, kann festgestellt werden, dass Google mit dem Wert 0,551 als Testsieger bestimmt wurde, während Ly-

cos mit 0,488, Fireball mit 0,391 und AltaVista mit 0,396 die hinteren Plätze einnahmen. Bei der systemorientierten Bewertung der Suchmaschinen konnte also Google als effektivste Suchmaschine herausgestellt werden. Nichtsdestotrotz sind die Werte der Microprecision aller Suchmaschinen an sich als nicht besonders hoch anzusehen: immerhin konnten sogar beim Erstplatzierten Google 45% der ausgegebenen Treffer bis zur Position 20 nicht als relevant oder auf ein relevantes Dokument verweisend, eingestuft werden. Bei der Ermittlung der Microprecision für die Retrievaleffektivität der Suchmaschinen bei einzelnen Suchanfragen, konnte folgendes festgestellt werden: bei den 56 gestellten Suchanfragen, konnte Google 32 mal den höchsten, 17 mal den zweithöchsten 7 mal den dritthöchsten und kein mal den niedrigsten Precisionwert aufweisen und erreichte somit den höchsten Microprecisionwert, gefolgt von Lycos, Fireball und AltaVista. Auch aus der nutzerbezogenen Sicht konnte Google somit als am effektivsten bezeichnet werden, während bei den anderen Suchmaschinen kein signifikanter Unterschied festgestellt werden konnte. Bei der Anzahl der insgesamt mit mindestens einem relevanten Treffer beantworteten Suchanfragen lag dagegen AltaVista vor Google. Große Unterschiede konnten beim Test mit den 4 Suchmaschinen allerdings nicht festgestellt werden (vgl. GRIESBAUM /RITTBERGER /BEKAVAC 2002, S. 14-20).

2004 wurde der Test mit einer ähnlichen Vorgehensweise wiederholt, wobei diesmal die Suchmaschinen Google.de, Lycos.de und AltaVista.de getestet worden sind. Was die Ergebnisse betrifft, konnte Google wieder das Beste erzielen, wobei die Unterschiede zwischen den 3 Suchmaschinen noch geringer ausgefallen aber insgesamt die Werte gestiegen sind. Bei der Anzahl der als relevant bewerteten Treffer aus ihrer Gesamtzahl von 1000, konnte Google 591 relevante Treffer aufweisen, Lycos 530 und AltaVista 510. Die Mean Average Precision (Micro) lag bei Google bei 0,65 bei Lycos erreichte sie 0,60 und bei AltaVista 0,56 sodass daraus geschlossen werden konnte, dass sich die Suchmaschinen in Hinblick auf die Precision stark ähneln (GRIESBAUM 2004). Auch die Ergebnisse aus der Macroperspektive bestätigten die Untersuchung von 2002: Von den 50

gestellten Suchaufgaben hatten in 24 Fällen die Suchergebnisse von Google die höchste Precision erreicht, gefolgt von Lycos mit 17 und AltaVista mit 13. (GRIESBAUM 2004)

### 3.3.2 Der Retrievaltest von Véronis

Die Ergebnisse von Griesbaum wurden von einer durch Véronis im Jahr 2006 durchgeführten Untersuchung bestätigt. Getestet wurde dabei die Qualität der drei amerikanischen Suchmaschinen Google, Yahoo und MSN sowie die drei französischen Suchmaschinen Exalead, Voilà und Dir.com. Als Juroren wurden 14 Studenten akquiriert, von denen jedem ein Thema (Tiere, Prominente, Kino, Aktuelles, Entertainment, Gesundheit, Geschichte, Literatur, Musik, Natur, Politik, Sport, Übernatürliches, Reisen) zugeteilt wurde. Zu diesem Thema sollte jeder Juror 5 freigeählte informationsorientierte Suchanfrage formulieren die anschließend an alle 6 Suchmaschinen geschickt wurden. Die ersten 10 Suchergebnisse jeder Suchanfrage wurden daraufhin gespeichert. Insgesamt ergaben sich aus den 14 Themen x 5 Suchaufgaben x 6 Suchmaschinen x 10 Suchtreffer 4.200 Webdokumente, von denen 3450 nach dem Ausschluss von Dopplungen zur Bewertung übrig geblieben sind. Es sollte bewertet werden, ob das Webdokument überhaupt zugänglich ist, es anstößige/ pornografische Inhalte enthält, zum Thema passt, es sich um eine E-Commerce Seite handelt und schließlich wie hoch seine Relevanz auf einer Skala von 0 (nicht relevant) bis 5 (sehr relevant) ist (vgl. VÉRONIS 2006, S.3).

Was die Relevanz betrifft, konnte festgestellt werden, dass insgesamt alle Suchmaschinen keine sehr hohen Werte erreicht haben und die Werte nah beieinander liegen. Google und Yahoo führten mit einer durchschnittlichen Relevanzbewertung von 2,3 gefolgt von MSN mit 2,0, Exalead mit 1,8, Dir mit 1,4 und Voilà mit 1,2. Es konnte somit geschlussfolgert werden, dass die Suchmaschinen ähnlich gut geeignet für die Beantwortung von informationsorientierten Suchanfragen sind (vgl. VÉRONIS 2006, S.7).

Auch wurden die Relevanzwerte in Abhängigkeit zu den Positionen der Treffer analysiert, die bis auf Dir und Voilà mit zunehmender Rankingposition abnahmen. So konnte festgestellt werden, dass das Rankingverfahren der Suchmaschinen Google, Yahoo, MSN und Excalead durchaus seine Richtigkeit hat (vgl. VÉRONIS 2006, S.8).

### **3.3.3 Der Retrievaltest von Lewandowski**

Ein Retrievaltest zur Überprüfung der Qualität der Suchmaschinen Google, Yahoo, MSN, Ask.com und Seekport hinsichtlich der ausgegebenen Webdokumente und der eigentlichen Treffer samt Titel, URL und Snippet wurde von Lewandowski im Jahr 2008 veröffentlicht. Ebenfalls standen informationsorientierte, reale Suchanfragen im Vordergrund der Untersuchung, die von den 40 Juroren zu gemischten Themen selbst formuliert wurden. Die Ausgegebenen Treffer und darauf verweisende Webdokumente wurden daraufhin anonymisiert und gemischt, sodass nicht nachvollziehbar war, aus welcher Suchmaschine sie stammen und auf welchem Rang sie ursprünglich gelistet waren. Dabei wurden den Juroren die ersten 20 Treffer jeder Suchmaschine auf die dazugehörige Suchanfrage zur binären Bewertung vorgelegt (vgl. LEWANDOWSKI 2008B, S.5).

Die Untersuchung ergab, dass 48,5% der von Yahoo ausgegebenen Webdokumente als relevant bewertet wurden, dicht gefolgt von Google mit 47,9%, wobei dieser Unterschied als nicht signifikant zu beurteilen ist. Bei der Bewertung der Treffer an sich konnte allerdings festgestellt werden, dass Google mit 60,2% als relevant beurteilten Treffer deutlich vor Yahoo mit 52,8% und den anderen Suchmaschinen liegt. Qualitätsunterschiede zwischen den Suchmaschinen hinsichtlich der Webdokumente konnten nicht eindeutig festgestellt werden, allerdings kann die Trefferdarstellung von Google, die vermehrt als relevant bewertet wurde, zur Begründung der Popularität der Suchmaschine herangezogen werden (vgl. LEWANDOWSKI 2008B, S.10-14).

Auch die Microprecision wurde ermittelt, um Unterschiede in der Relevanzbewertung nach Positionen feststellen zu können. Bei allen Suchmaschinen konnte eine fallende Tendenz der Relevanzbewertung mit zunehmender Rankingposition beobachtet werden, wobei Google und Yahoo insgesamt am besten abgeschnitten haben. Hinsichtlich der Effektivität der Suchmaschinen bezogen auf die einzelnen Suchanfragen wurde die Macroprecision hinzugezogen. Bei dieser Untersuchung konnte festgestellt werden, dass Yahoo bei 16 Suchanfragen diese am besten beantworten konnte wobei auch die Ergebnisse von Google überzeugen konnten (vgl. LEWANDOWSKI 2008B, S.10-14).

### **3.4 Erkenntnisse über Retrievaltests zur Suchmaschinenevaluierung**

Vergleicht man die vorgestellten Retrievaltests für informationsorientierte Suchanfragen von Griesbaum, Veronis und Lewandowski miteinander, kann man feststellen, dass sie sich von der Methodik und den ermittelten Ergebnissen stark ähneln (vgl. Tabelle 2). Größtenteils konnten Google und Yahoo die relevantesten Ergebnisse liefern und somit die höchste Retrievaleffektivität aufweisen. Was die Relevanzwerte hinsichtlich der einzelnen Trefferpositionen betrifft, wurde festgehalten, dass diese mit zunehmender Rankingposition sinken. Insgesamt waren diese, sowohl die Micro- als auch die Macroprecision, in den Tests nicht besonders hoch, was allerdings nicht nur an der Qualität der Suchmaschinen zu begründen ist, sondern auch an der Tatsache, dass für die Qualitätsmessung von Suchmaschinen in Retrievaltests Kennzahlen neu- und weiter entwickelt werden sollten.

|                           | <b>Griesbaum 2002/4</b>   | <b>Véronis 2006</b>  | <b>Lewandowski 2008</b>  |
|---------------------------|---|--|--|
| <b>Suchmaschinen</b>      | AltaVista, Fireball, Google, Lycos  | Google, Yahoo, MSN, Exalead, Voilà, Dir  | Google, Yahoo, MSN, Ask, Seekport  |
| <b>Suchanfragen</b>       | 56  | 70   |  |
| <b>Juroren</b>            | 28  | 14   | 40   |
| <b>Relevanz-Bewertung</b> | 3-Skala:<br>n/relevant/auf relevantes Dokument verweisend   | 5-Skala  | binär  |
| <b>Cut Off Wert</b>       | 20  | 10   | 20   |
| <b>Kennzahlen</b>         | Anzahl mit relevanten Treffern beantworteten Suchanfragen<br>Microprecision<br>Macroprecision                                 | Durchschnittlicher Relevanzwert,<br>Durchschnittlicher Relevanzwert/Position   | Anzahl mit relevanten Treffern beantworteten Suchanfragen<br>Microprecision<br>Macroprecision                                    |
| <b>Ergebnis</b>           | Google als Testsieger, Relevanz nimmt mit zunehmender Rankingposition ab, Unterschiede zwischen weiteren Suchmaschinen gering | Google und Yahoo führend, Relevanz nimmt mit zunehmender Rankingposition ab, Unterschiede zwischen weiteren Suchmaschinen gering | Google und Yahoo führend, Relevanz nimmt mit zunehmender Rankingposition ab, Unterschiede zwischen weiteren Suchmaschinen gering |

**Tabelle 2: Vergleich der Retrievaltests hinsichtlich Methodik und Ergebnisse**

Gleichzeitig wird sichtbar, dass die vorgestellten Tests, ähnlich wie zahlreiche weitere, sich auf einen Cut-Off-Wert von maximal 20 beschränken. Häufig standen auch die Ergebnisse lediglich aus den Top 3 Positionen im Vordergrund, da diese durch das tatsächliche Nutzerverhalten von besonderem Interesse sind. Es wird deutlich, dass dabei nicht unbedingt die Überprüfung der Relevanz der Suchergebnisse niedrig geranker Suchergebnisse im Vordergrund stand. Eine solche Untersuchung würde dabei neue Erkenntnisse bringen, vor allem hinsichtlich der Fragen, ob die Retrievaleffektivität der Suchmaschine bis zu den hinteren Positionen gegeben, ob das Ranking der Suchmaschine zuverlässig ist und ob das Verhalten der Nutzer, sich bei der Wahl der Treffer auf die ersten Positionen zu fixieren, berechtigt ist.

## **4 Relevanzanalyse niedrig geranker Google-Suchergebnisse**

Die Ergebnisse aus den Untersuchungen, sowohl aus der Nutzerforschung als auch der Evaluierung von Suchmaschinen durch Retrievaltests zeigen, dass eine Suchmaschinenevaluierung anhand einer Relevanzanalyse niedrig geranker Suchergebnisse auf informationsorientierte Suchanfragen, hier bei der Suchmaschine Google, durchaus einen Erkenntnisfortschritt für beide Forschungsbereiche darstellen würde. Ein solcher Evaluationstest soll den Kern dieser Bachelorarbeit ausmachen. Ziel ist, nicht etwa der Vergleich der Qualität verschiedener Suchmaschinen untereinander oder mit anderen Information Retrieval Systemen, sondern die Ermittlung der Retrievaleffektivität der Suchmaschine bis auf niedriger gerankte Positionen und eine Überprüfung des Rankings der Suchmaschine. Im Folgenden soll die grundsätzliche Vorgehensweise und das Studiendesign erläutert, die Testdurchführung beschrieben und die Ergebnisse wiedergegeben werden.

### **4.1 Studiendesign**

Die Vorgehensweise des durchgeführten Retrievaltests basiert hauptsächlich auf den Prinzipien der Cranfield- und TREC Methodik, bei denen Suchanfragen an die zu prüfende Suchmaschine abgeschickt, die ausgegebenen Webdokumente von Juroren auf ihre Relevanz hin bewertet und anschließend die Ergebnisse durch Kennzahlen wie Precision analysiert werden. Die Informationsbedürfnisse, Suchanfragen und der Bewertungsvorgang werden zuvor definiert und den Juroren vorgegeben, um die Vergleichbarkeit der Testergebnisse zu wahren. Die weiteren Kriterien für einen Retrievaltest, die speziell auf die Eigenschaften von Suchmaschinen ausgerichtet sind, wie etwa bei Hawking et al. und dem Framework von Lewandowski wurden ebenfalls hinzugezogen (vgl. Kapitel 3.2). Im nachfolgend beschriebenen Studiendesign wird ihre Anwendung verdeutlicht und die Inhalte und Eigenschaften des Tests erläutert.

#### **4.1.1 Wahl der Suchmaschine**

Da kein Vergleich zwischen Suchmaschinen oder Information Retrieval Systemen erfolgen soll, sondern einer von der Relevanz der Webdokumente auf verschiedenen Trefferpositionen lediglich einer Suchmaschine, wird die deutschsprachige Version von Google für den Test anhand der Popularität und dem intensiven Nutzungsgrad der Suchmaschine gewählt (vgl. WEB-STATS.INFO 2013, WEBHITS.DE 2013, COMSCORE.COM 2013).

#### **4.1.2 Verwendete Suchanfragen**

Um möglichst repräsentative informationsorientierte Suchanfragen und Informationsbedürfnisse abzubilden, die sowohl die thematische Vielzahl als auch die typische Formulierung dieser durch die Nutzer widerspiegeln, wurden reale Suchanfragen von Studenten des Departments Information an der Hochschule für Angewandte Wissenschaften Hamburg durch eine Befragung im Zeitraum vom 15. bis 19.7.2013 erhoben. Dazu wurden Fragebögen verteilt, die jeder Student mit seiner zuletzt formulierten informationsorientierten Suchanfrage und dem dahinter stehenden Informationsbedürfnis als Kontextinformation ausfüllen sollte. Dabei wurden 217 Suchanfragen mit den dazu passenden Hintergrundinformationen zur Suche ermittelt, von denen nach einer Überprüfung auf Dopplungen und auf den verlangten Suchanfragetyp, 200 übrig geblieben sind, die für den Retrievaltest verwendet werden.

#### **4.1.3 Anzahl der Suchergebnisse**

Um Pro Suchanfrage Webdokumente von den Toppositionen als auch vom möglichst niedrigen Rank in den Test einschließen zu können, wobei trotzdem eine überschaubare Menge an Webdokumenten den Juroren zur Bewertung vorgelegt werden sollte, werden die von der Suchmaschine Google ausgegebenen Webdokumente aus den Positionen 1-10, 13, 18, 23, 28, 33, 38, 43, 48, 51-60, 63, 68, 73, 78, 83, 88, 93 und 98 für den

Test verwendet. So können die Webdokumente der kompletten ersten Seite (1-10) und der kompletten sechsten Seite (51-60) miteinander verglichen und gleichzeitig ein Verlauf mit den dazwischen liegenden Einzelseiten im 3er und 8er Intervall bis hin zur 10. Seite (98) dargestellt werden. So ergeben sich pro Suchanfrage 36 Webdokumente, die zur Bewertung vorgelegt werden, wobei die Positionen der Webdokumente für den Juror anonymisiert und auch die Reihenfolge zufällig erfolgt, um Lerneffekte zu verhindern und eine unvoreingenommene Bewertung der Testpersonen zu erhalten. Die Trefferbeschreibungen mit Titel, Snippet und URL sollen in diesem Fall nicht berücksichtigt werden, da der Test ausschließlich der Relevanzbewertung anhand der Webdokumente an sich dienen soll.

#### **4.1.4 Die Juroren**

Von allen Studenten des Departments Information der HAW Hamburg die über den Test per E-Mail benachrichtigt wurden, nahmen die ersten 200 als Juroren daran teil. Jeder erhielt jeweils eine Suchanfrage mit der dazugehörigen Kontextinformation zum Informationsbedürfnis um sämtliche ausgegebene Webdokumente (also 36) eindeutig zu bewerten. Nach erfolgreicher Teilnahme erhielt jeder Juror einen Gutschein im Wert von 5,- der als Motivation zum Test dienen soll.

#### **4.1.5 Bewertung der Relevanz**

Obwohl der Begriff der Relevanz in seiner Definition nicht ganz eindeutig und problematisch durch die Subjektivität des Relevanzempfinden ist (vgl. Kapitel 3.1), steht er zur Bewertung der ausgegebenen Webdokumente in Retrievaltests von Suchmaschinen im Vordergrund, so auch in diesem. Die Relevanzbewertung wird von den Juroren anhand einer Skala von 1 bis 100 (1 nicht relevant, 100 sehr relevant) vorgenommen, um möglichst differenzierte und kalkulierbare Daten zu erhalten. Von einem binären Bewertungssystem wird abgewichen, um in der Untersuchung nicht nur nach relevanten und nicht relevanten Webdokumenten unterscheiden, sondern um besonders hoch relevante Dokumente ermitteln zu können. Die Be-

wertung wurde durch einen Regler ermöglicht, der von den Juroren nach links (nicht relevant) und rechts (sehr relevant) verschoben werden konnte, nachdem diese die Webdokumente kurz sichten und anschließend auf ihre Relevanz auf die gestellte Suchanfrage hin intuitiv bewerten sollten.

#### **4.1.6 Kennzahlen zur Datenauswertung**

Im nächsten Schritt soll eine statistische Auswertung der erhobenen Daten erfolgen, um diese zu ordnen und anschließend interpretieren zu können. Es wird von den einzelnen von Google ausgegebenen Webdokumenten als Merkmalsträger ausgegangen, die das Merkmal Position mit der Merkmalsausprägung des Rankingplatzes von 1-10, 13, 18, 23, 28, 33, 38, 43, 48, 51-60, 63, 68, 73, 78, 83, 88, 93 und 98 und das Merkmal Relevanz mit der Merkmalsausprägung der Relevanzbewertung von 1 bis 100 haben. Da auf eine klassische binäre oder Skalenbewertung der Relevanz, wie sie häufig in Retrievaltests verwendet wird (vgl. Kapitel 3.4) gezielt verzichtet und stattdessen eine Bewertung mittels Regler verwendet wurde, bei der die Abstände zwischen den einzelnen Werten stets gleich sind, ist eine differenzierte Analyse der Daten möglich. Es wird deshalb von einer reinen Auswertung mittels der populären Kennzahl Precision abgewichen, bei der die Webdokumente in relevante und nicht relevante unterteilt werden.

Die in diesem Test durchgeführte Datenanalyse soll Aussagen darüber ermöglichen, wie sich die Retrievaleffektivität der Suchmaschine mit zunehmender Rankingposition entwickelt, die Relevanz niedrig geranker Treffer im Vergleich zu den Toppositionen abschneidet, ob das Ranking der Suchmaschine demzufolge funktioniert und das Verhalten der Nutzer zu rechtfertigen ist. Dazu werden als erstes die für das Forschungsziel relevanten Fragestellungen des Tests durch einfache Häufigkeiten präsentiert, dann die Lagemaße zur Bestimmung der zentralen Tendenz der Häufigkeitsverteilung bezogen auf die Relevanzbewertungen ermittelt. Es werden folgende Kennzahlen benutzt:

Mittelwert, auch Durchschnitt, arithmetisches Mittel: beschreibt den statistische Mitte der Daten durch das Dividieren aller Werte des Datensatzes mit der Summe der Anzahl aller Werte (wenn  $n = \text{Grundgesamtheit, Menge}$  und  $x = \text{Merkmalsausprägungen, Werte}$ )

$$\bar{x} = \frac{(x_1 + x_2 + \dots + x_n)}{n}$$

Median, auch Zentralwert: stellt den mittig liegenden Wert der Datenverteilung dar, wenn diese in einer der Größe nach geordneten Reihe geordnet wurden, sodass die Hälfte der Daten immer kleiner, die andere größer als der Median sind (wenn  $n = \text{Skalenreihenwerte}$  ungerade)

$$Z = \frac{n + 1}{2}$$

Ist die Anzahl der Daten gerade, ist der Median die Hälfte der Summe der beiden in der Mitte liegenden Werte.

$$\frac{n}{2}$$

und Modus, auch Modalwert: gibt den am häufigsten besetzte Merkmalsausprägung der Befragung an. (vgl. RÖßLER/UNGERER 2008, S.8).

Die Kennzahlen werden dabei zum einen über alle Suchanfragen hinweg pro Position errechnet (vergleichbar mit der Mikromethode) für eine systemorientierte Bewertung der Suchmaschinenqualität und zum zweiten für alle Suchvorgänge gebildet (vergleichbar mit der Makromethode) um eine nutzerorientierte Bewertung der Suchmaschineneffektivität pro Suchanfrage zu erreichen. Für einen direkten Vergleich der Positionen 1 bis 10 als gut gerankte Treffer mit den Positionen 51 als 60 als niedrig gerankte Treffer sollen die jeweiligen Werte zusätzlich gegenübergestellt werden.

Anschließend werden Streuungsmaße zur Überprüfung der Streuung bzw. Disposition der durch die Lagemaße ermittelten Häufigkeitsverteilung herangezogen. So wird überprüft, ob etwa die errechneten Mittelwerte als

repräsentativ bezeichnet werden können, oder durch eine zu große Streuung der Werte ihre Aussagekraft verlieren. Dazu werden die Kennzahlen berechnet:

Standardabweichung: gibt die durchschnittliche Entfernung der gemessenen Ausprägungen eines Mittelwertes vom Durchschnitt an (wenn  $\bar{x}$  = Mittelwert)

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Varianz: Quadrat der Standardabweichung, kennzeichnet die Verteilung der Werte um den Mittelwert

$$v = s^2$$

Variationskoeffizient: gibt das Verhältnis in % zwischen Mittelwert und Standardabweichung an

$$V = \frac{s * 100}{\bar{x}}$$

Auch werden Maße zum Vergleich der Häufigkeitsverteilung mit der Normalverteilung verwendet, wie:

Schiefe: Neigungsstärke der Häufigkeitsverteilung nach links oder rechts im Vergleich mit Normalverteilung

$$\frac{n}{(n - 1)(n - 2)} \sum \left( \frac{x_i - \bar{x}}{s} \right)^3$$

Kurtosis: Wölbung der Häufigkeitsverteilung flach oder spitz im Vergleich zur Normalverteilung

$$w = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4$$

(vgl. RÖßLER/UNGERER 2008, S.10)

Im letzten Schritt sollen Zusammenhangsmaße verwendet werden, um die Abhängigkeit zwischen den beiden Variablen Position und Relevanz zu bestätigen oder zu widerlegen:

Korrelationsanalyse: misst den rechnerischen Zusammenhang zwischen zwei Merkmalen, hier für zwei metrische Merkmale mit der Kennzahl Korrelationskoeffizient nach Pearson, der einen Wert zwischen -1 (sehr geringer Zusammenhang) und 1 (sehr starker Zusammenhang) annehmen kann. ( $x_i$  und  $y_i$  Werte der Merkmale, Variable  $x$  und  $y$  jeweiliger Mittelwert,  $n$  Anzahl der Wertepaare)

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{(n - 1) * s_x * s_y}$$

(vgl. RÖßLER/UNGERER 2008, S.13, BÜHL 2012, S.346).

Regressionsanalyse: die einfache Regression stellt die Abhängigkeiten der Variablen fest (b Regressionskoeffizient, a Koordinaten der y-Achse). Der Einfluss der unabhängigen Variablen wird dabei mit dem quadrierten Koeffizienten  $R^2$  berechnet, dessen Wert zwischen 0 (unabhängig) und 1 (abhängig) liegt. Auch die Signifikanz muss durch den F-Test miteinbezogen werden, wobei ein hoher signifikanter Einfluss und Zusammenhang bei 0 und durch die Standard-Fehlertoleranz von 5% bei 0,05 liegt.

$$Y = b * x + a$$

(vgl. BÜHL 2012, S.434)

## 4.2 Umsetzung des Retrievaltests im RAT

Nachdem das Studiendesign festgelegt war, wurde der Retrievaltest im Relevance Assessment Tool als Online-Umfrage umgesetzt. Alle vier Teilkomponenten des Onlinetools, Suchmaschinen-Scraper, Backend, Frontend und das Auswertungsmodul wurden dafür verwendet. Im Folgenden werden die einzelnen Schritte in den jeweiligen Bereichen erläutert um die Entwicklung vom Studiendesign zum Retrievaltest aufzuzeigen.

#### 4.2.1 Verwendung des RAT Backend zum Anlegen des Testrahmens

Die Backend Funktion des RAT wurde zum Administrieren des Retrievaltests benutzt. Nach dem Aufrufen der Toolwebseite über einen Webbrowser und dem Einloggen als Administrator wurde ein neues Projekt für den Test angelegt. Dabei wurde ein Projektname „Relevanzanalyse niedrig geranker Suchergebnisse“ vergeben. Zusätzlich wurde eine Projektbeschreibung verfasst, die in der Testdurchführung den Juroren als Orientierung und Anleitung des Tests dient, sowie ein Abschlusstext, der den Juroren über die erfolgreiche Testteilnahme und den als Motivation dienenden Gutscheincode benachrichtigt. Die Projektsprache wurde so eingestellt, dass die Benutzerführung durch das jeweilige Template auf Deutsch erfolgte. Als zu testende Suchmaschine wurde google.de gewählt, als Anzahl der Suchanfragen pro Juror jeweils 1 definiert. Zudem wurde entschieden, dass jeder Nutzer sich mit demselben Zugangscod zum Test einloggen kann. Dieser wird in der E-Mail mit der Benachrichtigung und Aufforderung an der Testteilnahme den Studenten vermittelt.

Anschließend wurde ein neues Template für die Bewertungsskala erstellt, welche in dem Test verwendet wurde. Als Skalentyp wurde der Slider gewählt, als zu bewertende Elemente die Webdokumente festgelegt. Es wurde ein Text verfasst, der als Aufforderung an den Juror gerichtet war, die Relevanz der Webdokumente mit dem verschieben des Reglers nach links (relevant) oder rechts (nicht relevant) zu bewerten. Anschließend wurden die Werte für den Slider mit 1 als kleinster Wert und 100 als größter definiert und die Abstände zwischen den Werten in 1er Schritten gehalten. Dieses Template wurde anschließend in der Projektadministration als Bewertungsskala festgelegt.

Im nächsten Schritt wurde die Gutscheinfunktion des Tests konfiguriert. Dazu wurde ein vorgefertigtes Excel-Template heruntergeladen und mit den zugehörigen Gutscheincodes wieder dem System zugeführt. Nachdem zwei Texte, einer für den Fall einer erfolgreichen, der andere einer nicht

erfolgreichen Teilnahme verfasst wurden, wurde die Skiprate, also die Anzahl an tolerierten übersprungenen und somit nicht bewerteten Webdokumenten von 50 % festgelegt. Wurde diese überschritten, galt die Teilnahme am Test als nicht erfolgreich und somit wurde kein Gutscheincode verschickt.

Zuletzt wurden die für den Retrievaltest erhobenen Suchaufgaben hochgeladen. Dazu wurde ebenfalls ein vorgefertigtes Excel-Template runtergeladen, welches mit den Suchaufgaben und den dazugehörigen Suchbeschreibungen ausgefüllt und im System gespeichert wurde. Anschließend wurde festgelegt, dass sowohl die Suchaufgabe als auch die Kontextinformation im Test den Juroren angezeigt werden sollten.

#### **4.2.2 Erhebung der Webdokumente mit dem Suchmaschinen-Scraper**

Nachdem die Bedingungen des Tests definiert und dieser vollständig angelegt war, wurden in der weiteren Komponente des RAT, dem Suchmaschinen-Scraper, die Suchergebnisse von der Suchmaschine ermittelt, in das System kopiert und für die Verwendung im Test gespeichert. Dabei wurden zunächst die Scrapingintervalle festgelegt, gleich den zur Relevanz bewertenden Positionen 1-10, 13, 18, 23, 28, 33, 38, 43, 48, 51-60, 63, 68, 73, 78, 83, 88, 93 und 98. Bei dieser Prozedur wurden maximal 20 Suchanfragen auf einmal gescraped, um nicht von der Suchmaschine als Tool erkannt und ausgesperrt zu werden. Das Scrapen erfolgte insgesamt im Zeitraum vom 24. bis zum 31.7.2013. Nachdem die Webdokumente für die letzte der 200 Suchanfragen gescraped wurden, war der Retrievaltest vollständig angelegt und zum Durchführen mit den 200 Juroren bereit.

#### **4.2.3 Testdurchführung mit dem RAT Frontend**

Das Frontend bzw. dem Userinterface des Onlinetools dient der Testdurchführung und Datenerfassung der Relevanzbewertungen durch die Juroren. Wird die jeweilige URL im Webbrowser aufgerufen, kann die Test-

person nach Eingabe des Zugangscodes aus der Benachrichtigungs-E-Mail den Onlinetest starten.

Das Userinterface ist in drei Bereiche gegliedert. Im oberen Bereich befindet sich ein Fortschrittsbalken mit dem die Testpersonen die Dauer des Tests abschätzen können, die Suchaufgabe mit der dazugehörigen Aufgabenbeschreibung sowie der Regler zur Relevanzbewertung mit der Aufforderung die Relevanz der Webdokumente mit dem verschieben des Reglers zu bewerten. Im darunter liegendem Bereich wird das zu bewertende Webdokument angezeigt. Links davon befinden sich zwei Buttons, einer zum Überspringen des Webdokumentes im Falle einer nicht korrekten Darstellung, der andere zum Navigieren zum nächsten Webdokument, nachdem eine Bewertung abgegeben worden ist (vgl. Abbildung 4). Dabei werden pro Suchaufgabe alle 36 Webdokumente angezeigt, wobei ihre ursprüngliche Trefferposition anonymisiert wird und auch deren Reihenfolge zufällig ist. So soll das Frontend den Juroren die Bearbeitung der Suchaufgaben durch die Bewertung der ausgegebenen Webdokumente mittels Regler ermöglichen.

Bevor der so angelegte Test durchgeführt wurde, sollte durch einen Pretest überprüft werden, inwieweit die formulierten Aufforderungen und das Prinzip der Relevanzbewertung durch die Juroren angenommen und umgesetzt werden. Auch sollte geklärt werden, mit welcher Dauer für die Onlineumfrage gerechnet werden muss und ob so die geplante Bewertung von 36 Webdokumenten pro Juror in einem angemessenen Zeitrahmen erfolgen kann. Dazu nahmen 5 Studenten vor Erhebungsbeginn als Juroren an dem Test teil. Dabei wurde festgestellt, dass der Sinn und Zweck des Tests verstanden, die Relevanzbewertung keine Probleme darstellte und die Form und der Onlineumfrage positiv bewertet wurde. Auch konnte durch den Pretest ermittelt werden, dass mit einer Dauer von 10 bis 15 Minuten die Bewertung von 36 Webdokumenten pro Juror durchaus realistisch ist.

Am Testdesign des finalen Retrievaltest mussten nach den Ergebnissen des Pretests keine Veränderungen vorgenommen werden, sodass dieser am 1.8.2013 aktiviert und mit einer E-Mail an die Studenten des Departments angekündigt wurde. Bis zum 6.8.2013 wurden alle 200 Suchaufgaben durch die Juroren bewertet und so die Relevanzurteile ermittelt und automatisch durch das RAT gespeichert.

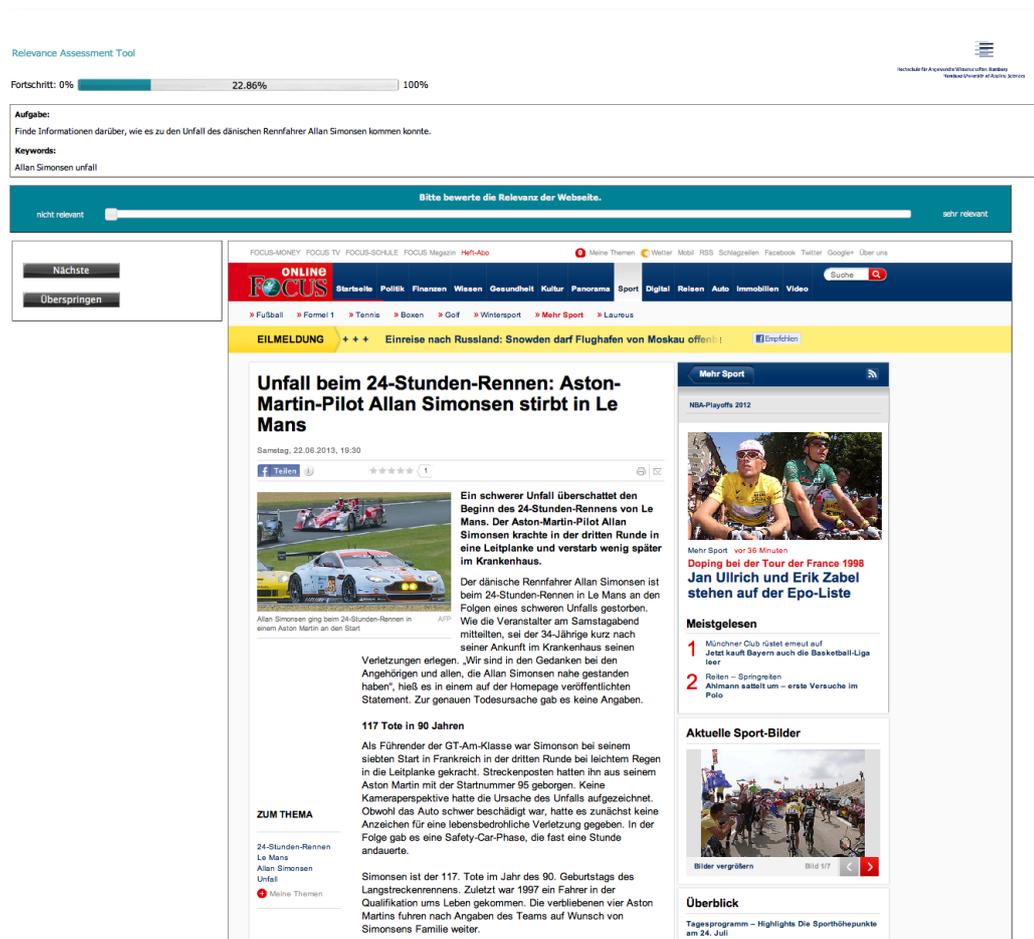


Abbildung 4: Darstellung des Userinterface anhand einer beispielhaften Suchaufgabe

#### 4.2.4 Aufbereitung der Daten mit dem RAT Auswertungsmodul

Die zuvor anonymisierten und gemischten Webdokumente wurden im Auswertungsmodul des RAT mit den durch die Juroren abgegebenen Relevanzbewertungen ihrem ursprünglichen Rankingpositionen automatisch zugeordnet. So konnten die Ergebnisse des Tests in Form einer Exceldatei exportiert werden, die die ID, URL, die Position und die Relevanzbe-

wertung der einzelnen Webdokumente, die ID der Juroren und der Wortlaut der 200 Suchaufgaben enthalten. Anhand dieser erhobenen Daten erfolgte anschließend die statistische Auswertung der Testergebnisse.

### 4.3 Ergebnisse der Studie

Durch die Auswertung des ermittelten Datensatzes mit dem Statistikprogramm SPSS können Aussagen über die Tendenz der Häufigkeiten der Relevanzwerte, wie auch ihre Verteilung und ihre Abhängigkeit mit der Position der Webdokumente gemacht werden. Die Ergebnisse der Auswertung sollen im Folgenden dargestellt werden.

#### 4.3.1 Häufigkeiten der Relevanzbewertung für alle Positionen

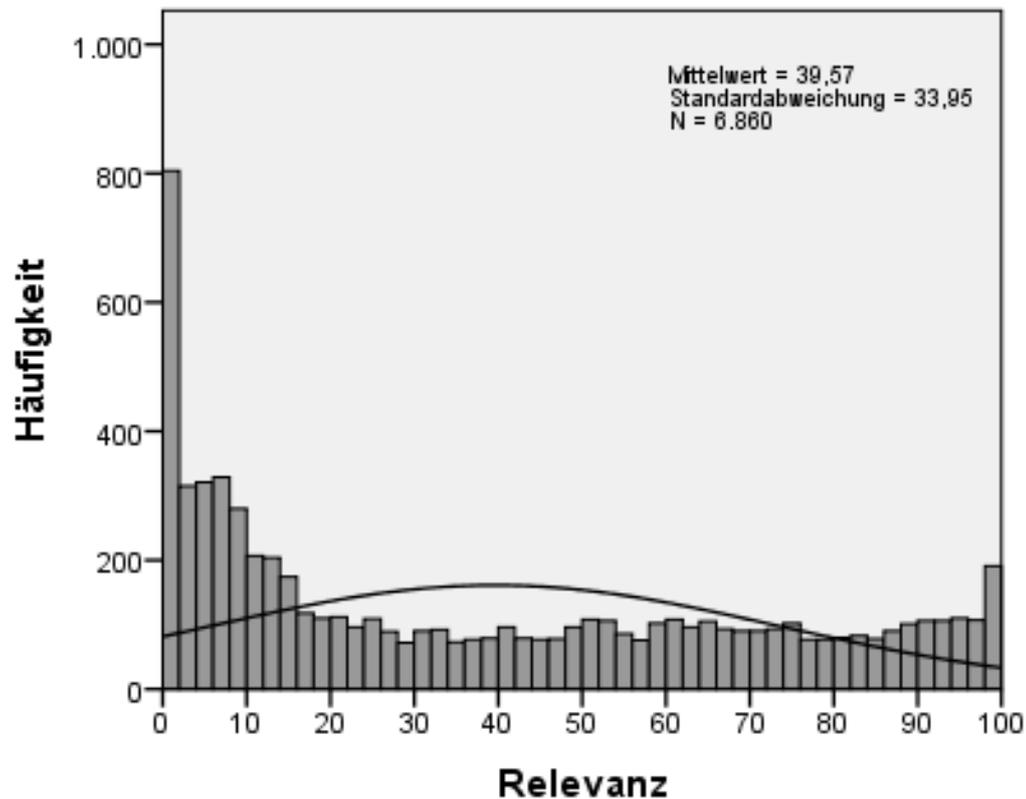
Die Grundgesamtheit des Datensatzes beträgt 7.182 Webdokumente, die im Retrievaltest den Juroren zur Bewertung vorgelegt wurden. Davon sind 322 Webdokumente (4,5%) durch Überspringen oder Auslassen nicht in die Auswertung einbezogen wurden, sodass insgesamt 6.860 gültige Bewertungen zustande gekommen sind (vgl. Tabelle 3).

|                    |         |          |
|--------------------|---------|----------|
| N                  | Gültig  | 6860     |
|                    | Fehlend | 322      |
| Mittelwert         |         | 39,57    |
| Median             |         | 31,50    |
| Modalwert          |         | 1        |
| Standardabweichung |         | 33,950   |
| Varianz            |         | 1152,596 |
| Schiefte           |         | ,418     |
| Kurtosis           |         | -1,286   |

**Tabelle 3: Häufigkeiten aller Positionen**

Über alle Positionen hinweg wurden die Webdokumente durchschnittlich mit einer Relevanz von 39,57 bewertet, wobei die hohen Werte der Streuungsmaße Standardabweichung (33,95), Varianz (1.152,59) und Variationskoeffizient (85,79%) eine sehr starke Streuung aufzeigen und somit eine hohe Abweichung der einzelnen Werte vom Mittelwert belegen. Im Vergleich zur Normalverteilung zeigen die Werte der Schiefe mit 0,418

(größer als 0) eine linksgipflige Häufigkeitsverteilung, die mit einem Kurtosiswert von  $-1,28$  (kleiner als 0) flacher verläuft (vgl. Abbildung 5).



**Abbildung 5: Histogramm zur Häufigkeitsverteilung der Relevanzwerte über alle Positionen und Suchanfragen hinweg**

11,7% von den 6.860 zur Bewertung vorgelegten Webdokumenten (804) erhielten den Relevanzwert von 1, sodass dieser am häufigsten vergeben und auch den Modus des Datensatzes ausmacht. Die relative Summenhäufigkeit von 38,4% zeigt, dass über  $1/3$  der Webdokumente mit einem Relevanzwert von 15 oder schlechter bewertet wurden. Dem gegenübergestellt wurde der Wert 100 insgesamt 222 Webdokumenten, also in 3,1% der Fälle zugeordnet, mit den ebenfalls hohen Werten 99 und 98 in für 5,7% der Webdokumente verwendet. Die dazwischen liegenden Werte von 16 bis 97 wurden dagegen mit jeweils höchstens 0,9% gleichermaßen selten und gleichmäßig auf die Webdokumente verteilt.

### 4.3.2 Häufigkeiten der Relevanzbewertung pro Position

Durch die Analyse der Relevanzbewertungen pro Position wird die Häufigkeitsverteilung mit den zuvor verwendeten Maßen über alle Suchanfragen hinweg für jede Position berechnet. Um die Ergebnisse der Relevanzbewertungen aller Position zusammenzufassen und vergleichbar zu machen, wird der Mittelwert der Relevanzbewertungen pro Position im Diagramm dargestellt (vgl. Abbildung 6), sodass eine Auswertung der Daten angelehnt an die Mikromethode zur systemorientierten Bewertung der Suchmaschinenqualität möglich ist. Demnach erhalten über alle Suchanfragen hinweg die Webdokumente der Position 1 mit einem Wert von durchschnittlich 58,98 (bzw. Mikroprecision 0,58) die höchsten Relevanzbewertungen. Diese nehmen mit zunehmender Rankingposition deutlich ab, sodass etwa die Webdokumente mit der Position 98 im Durchschnitt einen Relevanzwert von 29,68 (bzw. Mikroprecision 0,29) erreichen.

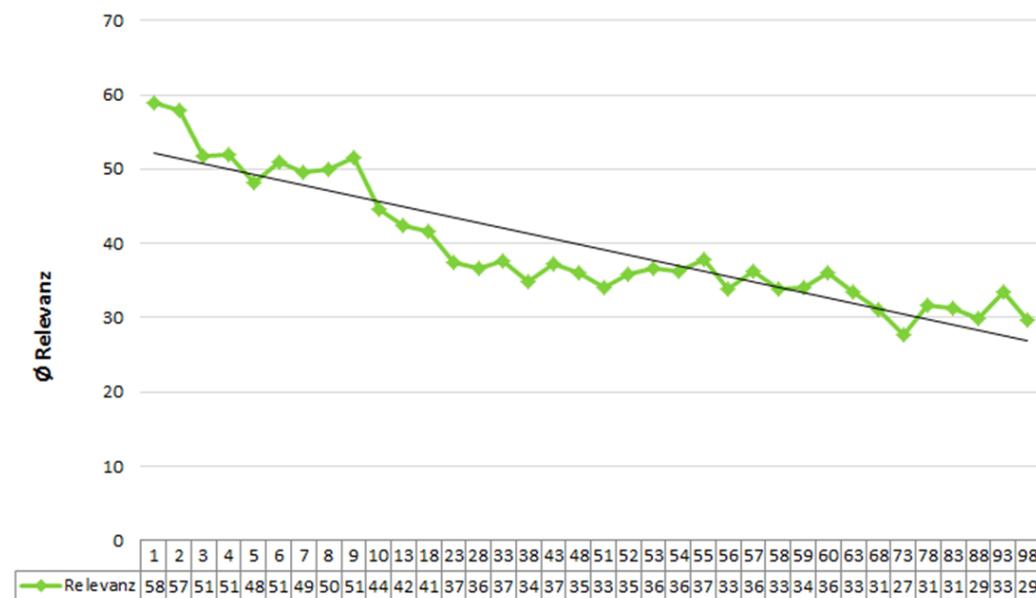


Abbildung 6: Durchschnittliche Relevanzbewertung pro Position

Stellt man das Lagemaß Median vergleichbar dar, welches durch die hoch ausfallenden Streuungsmaße ergänzend zur statistischen Auswertung herangezogen wurde, ergibt sich eine ähnliche Verteilung. Den höchsten Wert erreicht mit 69 ebenfalls die Position 1., den niedrigsten die Position

73 mit 12. Der Abwärtstrend in der Relevanz der Webdokumente mit zunehmender Position kann demnach bestätigt werden (vgl. Abbildung 7).

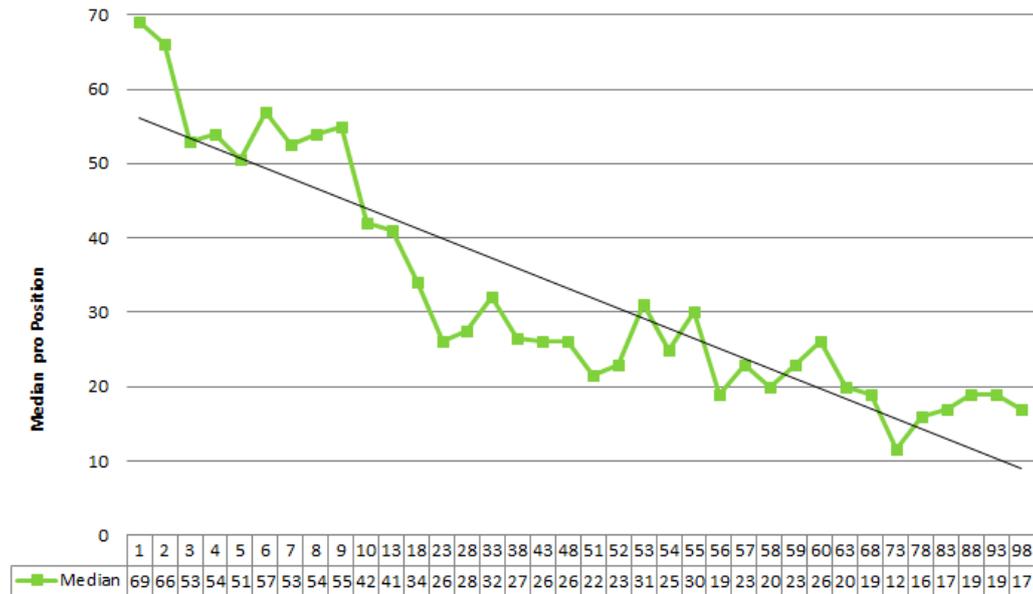
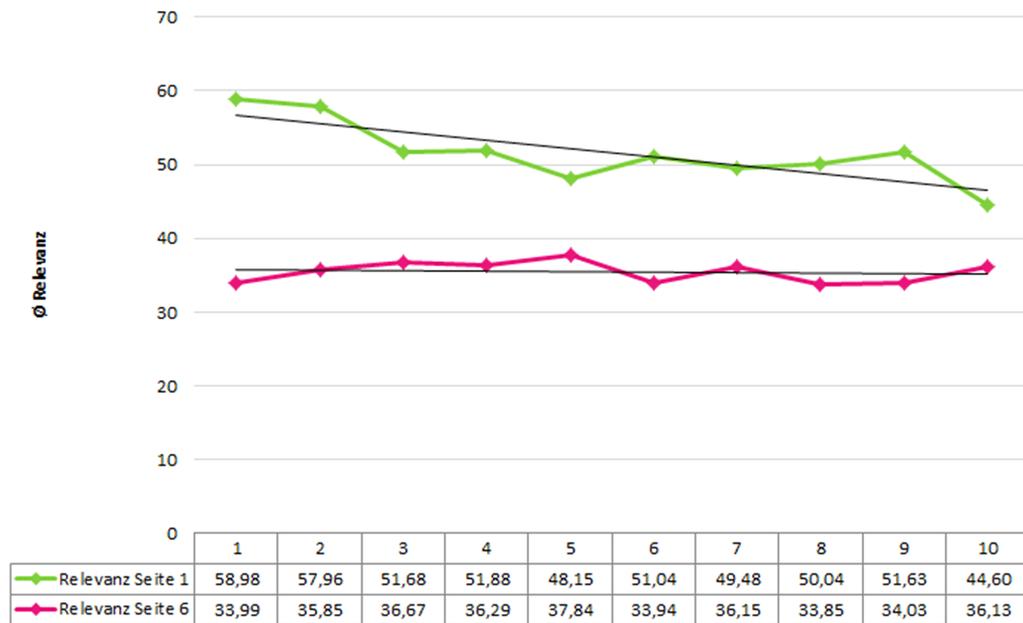


Abbildung 7: Median der Relevanzbewertung pro Position

### 4.3.3 Häufigkeiten der Relevanzbewertung der Seite 1 und 6 im Vergleich

Um Aussagen über die Relevanz gut geranker mit niedriger gerankten Suchergebnissen machen zu können, werden die durchschnittlichen Relevanzbewertungen der Treffer 1 bis 10 der Seite 1 mit der Seite 6 verglichen. Während die Treffer auf Seite 1 über alle Suchanfragen hinweg insgesamt eine durchschnittliche Relevanzbewertung von 51,52 (bzw. Microprecision 0,51) aufweisen, liegt dieser Wert der Treffer der Seite 6 bei 35,47 (bzw. Microprecision 0,35). Betrachtet man die durchschnittlichen Relevanzbewertungen der einzelnen Positionen kann festgestellt werden, dass mit einem Relevanzwert von 58,98 (bzw. Microprecision von 0,58) auf Position 1 und 44,60 (bzw. Microprecision 0,44) auf Position 10 die Relevanz auf Seite 1 mit zunehmender Rankingposition abnimmt. Dem gegenübergestellt kann dieser Abwärtstrend der Relevanz auf Seite 6 mit einem Relevanzwert der Position 1 von 33,99 (bzw. Microprecision 0,33) und 36,13 (bzw. Microprecision 0,36) auf Position 10 nicht festgestellt

werden. Hier bleiben die durchschnittlichen Relevanzwerte eher konstant (vgl. Abbildung 8).



**Abbildung 8: Durchschnittliche Relevanzbewertung pro Position der Seite 1 und 6**

Vergleicht man den Median der Relevanzbewertungen der Treffer 1 bis 10 der Seite 1 mit der Seite 6 werden ebenfalls Parallelen deutlich. Der Median des 1. Treffers auf Seite 1 erreicht einen Wert von 69 und sinkt auf Position 10 deutlich auf 42. Die Positionen auf Seite 6 bleiben mit Werten von 19 bis 31 eher konstant (vgl. Abbildung 9).



Abbildung 9: Median der Relevanzbewertung pro Position der Seite 1 und 6

#### 4.3.4 Häufigkeiten der Relevanzbewertung pro Suchanfrage

Um die Retrievaleffektivität bei der Beantwortung der einzelnen Suchanfragen für eine benutzerorientierte Qualitätsbewertung zu überprüfen (vergleichbar mit der Macromethode), wird ermittelt, bei wie vielen von den 200 insgesamt gestellten Suchanfragen die Seite 1 höhere Relevanzwerte erreicht hat und bei wie vielen die Seite 6 relevantere Ergebnisse lieferte. Abbildung 8 zeigt, dass auf allen Positionen die Suchaufgaben mit den Ergebnissen aus der Seite 1 mit relevanteren Ergebnissen beantwortet werden konnten als auf der Seite 6. Auf Position 1 erreichten die Webdokumente auf der Seite 1 bei 136 Suchanfragen höhere Relevanzwerte als bei Seite 6 – dort konnten lediglich 55 Suchanfragen besser von den Ergebnissen der Seite 6 beantwortet werden. Insgesamt nimmt mit zunehmender Position die Anzahl der durch die Seite 1 besser beantworteten Suchaufgaben ab und die auf der Seite 6 zu, nichtsdestotrotz bleibt erstere stets höher. Auf Position 10 kann diese Entwicklung aufgezeigt werden: zwar bleibt die Anzahl der durch relevantere Ergebnisse beantworteten Suchanfragen auf Seite 1 höher, allerdings ist die Differenz zwischen den beiden Werten (100 für Seite 1, 83 für Seite 6) bei weitem nicht so groß wie es für die anderen Positionen der Fall war (vgl. Abbildung 8).

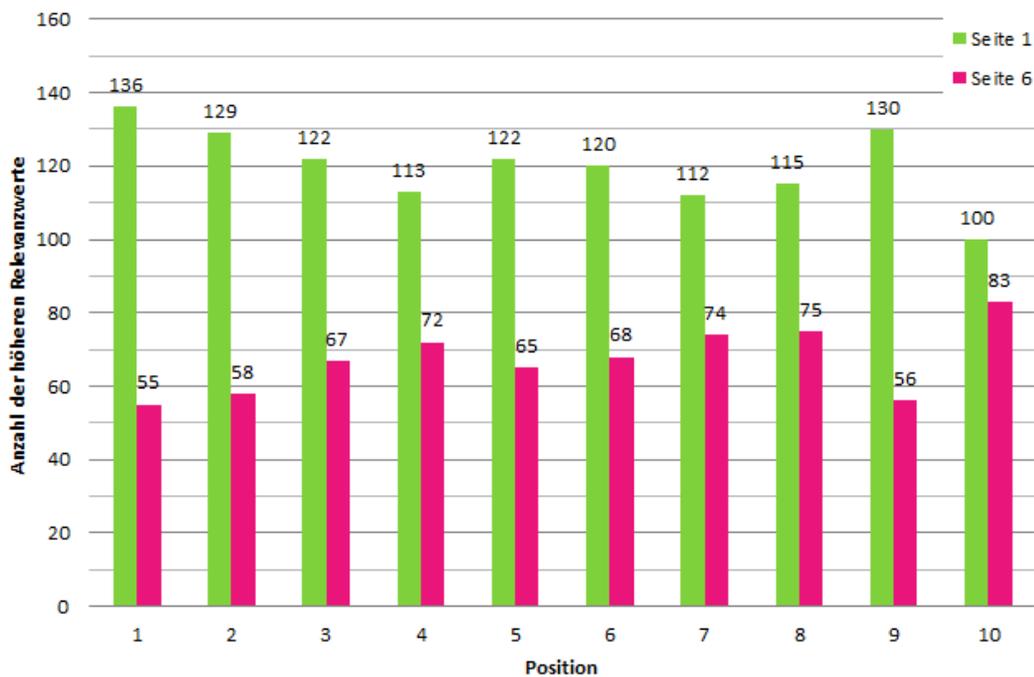


Abbildung 10: Anzahl höherer Relevanzwerte bei Seite 1 und 6 pro Position

### 4.3.5 Zusammenhang zwischen Position und Relevanz

Die dargestellten Häufigkeiten aus dem Test und ihre Verteilung, sowohl über alle Suchanfragen und Positionen hinweg als auch für diese einzelnen Variablen, geben eine Übersicht über die Merkmalsausprägungen. Um eine mathematische Abhängigkeit zwischen den Merkmalen Relevanz und Position zu überprüfen, wird eine Korrelations- und Regressionsanalyse der Daten durchgeführt und im Folgenden beschrieben.

Der Korrelationskoeffizient mit dem Wert von  $-0,218$  zeigt einen sehr geringen Zusammenhang zwischen den Merkmalen, wobei mit einem Wert von  $p = 0$  das Ergebnis sehr signifikant ist (vgl. Tabelle 4). Allerdings ist zu berücksichtigen, dass vor allem die Variable Relevanz als subjektive Wahrnehmung nicht als „hard fact“ bezeichnet werden kann, bei denen die Ergebnisse der Analyse eindeutiger auszulegen sind.

Korrelationen

|          |                     | Position | Relevanz       |
|----------|---------------------|----------|----------------|
| Position | Pearson-Korrelation | 1        | <b>-,218**</b> |
|          | Sig. (2-seitig)     |          | ,000           |
|          | N                   | 7182     | 6860           |

|          |                     |         |      |
|----------|---------------------|---------|------|
| Relevanz | Pearson-Korrelation | -,218** | 1    |
|          | Sig. (2-seitig)     | ,000    |      |
|          | N                   | 6860    | 6860 |

\*\* Korrelation ist bei Niveau 0,01 signifikant (zweiseitig).

**Tabelle 4: Korrelationsanalyse der Daten aus dem Retrievaltest**

Auch die Regressionsanalyse führt zunächst zu ähnlichen Ergebnissen mit einem Regressionskoeffizienten b von -0,256 und der Signifikanz von 0. Der R-Quadrat Wert von 0,48 bzw. 48% zeigt dagegen einen mittleren Einfluss der Position als unabhängige Variable auf die Relevanz der Webdokumente. Auch die Tatsache, dass der Wert R-Quadrat niedriger ist als die Irrtumswahrscheinlichkeit von 0,05 zeigt, dass die Variable x abhängig von y ist. Somit kann die Hypothese, dass die Relevanz der getesteten Webdokumente abhängig von der Position dieser ist, bestätigt werden (vgl. Tabelle 5, 6).

**Koeffizienten<sup>a</sup>**

|          | Nicht standardisierte Koeffizienten |                | Standardisierte Koeffizienten | t       | Sig. | Konfidenzintervall für B (95,0%) |            |
|----------|-------------------------------------|----------------|-------------------------------|---------|------|----------------------------------|------------|
|          | B                                   | Standardfehler | Beta                          |         |      | Untergrenze                      | Obergrenze |
|          | (Konstante)                         | 50,144         | ,697                          |         |      |                                  | 71,944     |
| Position | -,256                               | ,014           | -,218                         | -18,526 | ,000 | -,283                            | -,228      |

a. Abhängige Variable: Relevanz

**Tabelle 5: Regressionsanalyse der Daten aus dem Retrievaltest**

**Modellübersicht**

| Modell | R                 | R-Quadrat | Angepasstes R-Quadrat | Standardfehler der Schätzung |
|--------|-------------------|-----------|-----------------------|------------------------------|
| 1      | ,218 <sup>a</sup> | ,048      | ,048                  | 33,133                       |

a. Prädiktoren: (Konstante), Position

**Tabelle 6: R-Quadrat Test der Daten aus dem Retrievaltest**

## **4.4 Ergebnisse der Datenanalyse**

Durch statistische Auswertung der Daten aus dem Retrievaltest niedrig geranker Google-Suchergebnisse auf informationsorientierte Suchanfragen können einige Aussagen über die Retrievaleffektivität und das Relevanzranking der Suchmaschine Google gemacht werden. Diese werden im Folgenden, aufgeteilt auf die einzelnen zuvor erstellten Hypothesen erläutert.

### **4.4.1 Retrievaleffektivität der Suchmaschine bei den hinteren Positionen**

Die Hypothese, dass die Retrievaleffektivität bei informationsorientierten Suchanfragen von Google auch bei den hinteren Trefferpositionen gegeben ist, konnte durch die Ergebnisse des Tests nicht bestätigt werden. Die errechneten Mittel- und Medianwerte haben gezeigt, dass Webdokumente hinterer Trefferpositionen mit niedrigeren Relevanzwerten bewertet wurden und somit meist nicht in der Lage sind zufriedenstellende Ergebnisse auf die gestellten Suchanfragen zu liefern. Insgesamt war die Bereitschaft der Nutzer, hohe Relevanzwerte zu vergeben geringer. Nichtsdestotrotz zeigen die hohen Streuungsmaße, dass die Relevanzbewertungen pro Nutzer recht unterschiedlich getroffen worden sind. Die Bewertung eines Webdokuments lässt sich somit als eine individuelle Entscheidung der Nutzer festhalten, die subjektiv getroffen und somit unterschiedlich ausfallen kann.

### **4.4.2 Beurteilung des Relevanzrankings**

Die Hypothese, dass die Relevanzbewertung der Webdokumente durch die Nutzer mit fallender Rankingposition ebenfalls sinkt, konnte dagegen durch die errechneten Lagemaße bestätigt werden. Während den Treffern der Top-Positionen im Durchschnitt eine hohe Relevanz zugesprochen wurde, nahm diese mit niedrigerem Ranking kontinuierlich ab. Allerdings ist festzuhalten, dass dieser Abwärtstrend beim Vergleich mittig positionierter Treffer mit den am niedrigsten positionierten nicht mehr so stark auftritt. Die Relevanz der Webdokumente dieser Positionen kann also als

recht konstant bewertet werden. Besonders in Hinblick auf die Top-Positionen kann man von einem zuverlässigen Ranking der Webdokumente durch die Suchmaschine sprechen, während bei den mittleren und hinteren Positionen keine signifikanten Relevanzunterschiede gemacht werden und somit auch das Ranking. Vergleicht man speziell die Ergebnisse der Seite 1 mit denen der Seite 6, wird deutlich, dass erstere relevanter und somit effektiver die Suchanfragen beantworten. Bezogen auf das Nutzerverhalten der User bei der Evaluation der Suchergebnisse, kann von einem effektiven Rankingverfahren gesprochen werden. Dass sich die Nutzer bei der Sichtung der Treffer meistens auf die erste SERP beschränken, kann deshalb als intuitiv sinnvoll getroffenes Verhalten bewertet werden.

#### **4.4.3 Zusammenhang zwischen Relevanz und Position**

Mit der Bestätigung der vorangegangenen Hypothese konnte zudem ermittelt werden, dass ein Zusammenhang zwischen der Relevanz und der Position eines Webdokumentes besteht. Die Korrelations- und Regressionsanalyse haben gezeigt, dass die Relevanz eines Webdokuments von der Position abhängig ist.

## 5 Fazit und Ausblick

Der Retrievaltest im Rahmen der Relevanzanalyse niedrig geranker Google Suchergebnisse auf informationsorientierte Suchanfragen mit dem Relevance Assessment Tool hat gezeigt, dass die Retrievaleffektivität der Suchmaschine bei gut gerankten Treffern höher ist, als bei den darauffolgenden. Somit kann von einem effektiven Rankingverfahren der Suchmaschine gesprochen werden. Besonders ist dieses auf den Top Positionen gegeben, die bei der Evaluierung der Treffer durch die Nutzer im Fokus stehen. Die recht konstant bleibende Relevanz der mittleren und hinteren Positionen zeigt, dass das Ranking in dem Bereich nicht dieselbe Bedeutung hat. Da die Evaluation durch Nutzer diesen nicht einbezieht, sind Optimierungsmaßnahmen dafür nur bedingt sinnvoll. Allerdings erreichen auch die Top-Positionen durchschnittlich Relevanzwerte von oberem Mittelbereich, sodass sich die Frage nach dem Finden hochrelevanter Webdokumente durch die Suchmaschine stellt.

Dass Webdokumente vorderer Positionen als relevanter bewertet wurden, wurde durch den Test festgestellt. Von welchen Faktoren sich die Nutzer dabei leiten ließen, bleibt in dem Test unbeantwortet. Die Begründung der Relevanzbewertung der Webdokumente, etwa anhand der 7 Faktoren nach Giesbergen, Moost und Aelen könnte in zukünftigen Studien die Frage klären, warum ein Webdokument als relevant wahrgenommen wird.

Die Testdurchführung an sich zeigt, dass die Nutzung des RAT zum Aufbereiten, Durchführen und Auswerten von Retrievaltests eine aufwand- und zeitsparende Methode ist. Insbesondere die automatisierten Vorgänge beim Scraping, Suchaufgaben und der Umsetzung im Userinterface und der anschließenden Wiederzuordnung der Relevanzbewertungen zu den jeweiligen Positionen kann als solche angesehen werden.

## Literatur

### BRODER 2002

Broder, A.: *A taxonomy of web search*. SIGIR Forum 36 (2). [online]. – URL: <http://www.acm.org/sigir/forum/F2002/broder.pdf> (Abruf: 2013-07-18)

### BROPHY/BAWDEN 2005

Brophy, J; Bawden, D.: *Is Google enough? Comparison of an internet search engine with academic library resources*. Aslib Proceedings 57, 2005, 498-512.

### BÜHL 2012

Bühl, A.: *SPSS 16: Einführung in die moderne Datenanalyse*. 11 Aufl. München: Pearson Education Deutschland, 2012.

### CLEVERDON/MILLS/KEEN 1966

Cleverdon, C. W., J. Mills, and M. Keen. 1966. *Factors determining the performance of indexing systems, vol. 2: Test results. Technical report*. Aslib Cranfield Research Project, Cranfield, England.

### COMSCORE.COM 2013

comScore Releases June 2013 U.S. *Search Engine Rankings* [online]. – URL: [http://www.comscore.com/ger/Insights/Press\\_Releases/2013/7/comScore\\_Releases\\_June\\_2013\\_U.S.\\_Search\\_Engine\\_Rankings](http://www.comscore.com/ger/Insights/Press_Releases/2013/7/comScore_Releases_June_2013_U.S._Search_Engine_Rankings) (Abruf: 2013-07-18)

### CUTRELL/GUAN 2007

Cutrell, E.; Guan, Z.: *What are you looking for? An Eye-Tracking Study of Information Usage in Web Search*. CHI '07 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM New York, 2007 S.407

## ENQUIRO 2005

Enquiro: *Google Eye Tracking Report: How Searchers See and Click on Google Search Results*. Enquiro Search Solutions, 2005

## ENQUIRO 2006

Enquiro: *Eye Tracking Report: Google, MSN and Yahoo! Compared : An in depth look at interactions with Google, MSN & Yahoo! Using Eye Tracking Methodology*. Enquiro Search Solutions, 2006

## ENQUIRO 2008

Enquiro: *Barries on a Search Results Page*. Enquiro Search Solutions, 2008

## FAULDRATH 2011

Fauldrath, Jens: *SERP CTR Studien kritisch betrachtet*. 2011 [online] – URL: <http://www.inhouse-seo.de/serp-ctr-studien/> (Abruf: 2013-08-13)

## GRIESBAUM 2000

Griesbaum, J.: *Evaluierung hybrider Suchsysteme im WWW*. Konstanz, Universität Konstanz, Informationswissenschaft, Dipl.-Arbeit, 2000

## GRIESBAUM 2004

Griesbaum, J.: *Evaluation of three German Search Engines: Altavista.de, Google.de and Lycos.de*. Information Research 9, 2004 [online] – URL: <http://eprints.rclis.org/5746/1/paper189.html> (Abruf: 2013-08-13)

## GRIESBAUM/RITTBERGER/BEKAVAC 2002

Griesbaum, J.; Rittberger, M.; Bekavac, B.: *Deutsche Suchmaschinen im Vergleich: AltaVista.de, Fireball.de, Google.de und Lycos.de*. in: Hammwöhner, R.; Wolff, C.; Womser-Hacker, C. (Hrsg), Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information, Procee-

dings 8. Internationales Symposium für Informationswissenschaft, Konstanz 2002, S.201-223.

GISBERGEN/ MOST/AELEN 2007

Gisbergen, M.S. van; Most, J. van der; Aelen, P. (2007) *Visual attention to Online Search Engine*. [online] – URL: [http://www.akademor.webd.pl/download/eyetracking\\_research.pdf](http://www.akademor.webd.pl/download/eyetracking_research.pdf) (Abruf: 2013-07-18)

GONZÁLEZ-CARO/MARCOS 2010

González-Caro, C.; Marcos, M.: *Different Users and Intents: An Eye-tracking Analysis of Web Search*. Barcelona: Pompeu Fabra University, 2010

GOOGLE MARKTFORSCHUNG GERMANY 2007

Google Marktforschung Germany: *Betrachtungsverlauf einer SERP. Eine Eyetracking- Studie zur Google-Ergebnisseite* [online] – URL: <http://www.full-value-of-search.de> (Abruf: 2013-07-18)

GORDON / PATHAK 1999

Gordon, M.; Pathak, P.: *Finding information on the World Wide Web : the retrieval effectiveness of search engines*. In: *Information Processing and Management* 35, 1999, S. 141 – 180.

GRANKA et al. 2004

Granka, L.; Joachims, T.; Gay, G.: *Eye-Tracking Analysis of User Behavior in WWW Search*. Cornell University, 2004

GRANKA/FEUSNER/LORIGO 2008

Granka, L.; Feusner, M.; Lorigo, L.: *Eyetracking in Online Search*. In: *Passive Eye Monitoring*, 2008, S. 283-304.

HAWKING ET AL. 2001

Hawking D. ; Craswell, N.; Bailey, P.; Griffiths, K.: *Measuring Search Engine Quality*. In: Information Retrieval 1(4), 2001, S. 33 – 59.

HÖCHSTÖTTER 2007

Höchstötter, N.: *Suchverhalten im Web – Erhebung, Analyse und Möglichkeiten*. In: Information: Wissenschaft & Praxis, 58(2007)3, Karlsruhe: 2007

HÖCHSTÖTTER 2009

Höchstötter, N.: *Methoden der Erhebung von Nutzerdaten und ihre Anwendung in der Suchmaschinenforschung*. In: Lewandowski, D. (Hrsg.): Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis. Heidelberg : Akademische Verlagsgesellschaft Aka GmbH, 2009, S. 175 – 203

HÖCHSTÖTTER/KOCH 2009

Höchstötter, N.: Koch, M.: *Standard parameters for searching behaviour in search engines and their empirical evaluation*. Journal of Information Science, 35, 45-65, 2009

JANSEN/SPINK 2006

Jansen, B.J; Spink, A.: *How are we searching the World Wide Web? A comparison of nine search engine transaction logs*. Information Processing & Management 42 (2006), S. 248–263.

KEANE/O'BRIEN/SMYTH 2008

Keane, M.T.; O'Brien, M.; Smyth, B.: *Are people biased in their use of search engines?*. Communications of the ACM 51, 2008 S. 49-52.

LADSTÄTTER 2009

Ladstätter, M.: *Wie werden die AdWords auf <http://google.at> wahrgenommen? Ein Vergleich von Benutzern mit Kauf- und Informationsabsicht an-*

*hand von Eyetracking*. [online] – URL: <http://www.usability.at/> (Abruf: 2013-08-13)

LEWANDOWSKI 2005

Lewandowski, D.: *Web Information Retrieval. Technologien zur Informationssuche im Internet*. Frankfurt am Main: DGI, 2005 S.34

LEWANDOWSKI 2006

Lewandowski, D.: *Themen und Typen der Suchanfragen an deutsche Web-Suchmaschinen*. In: Lehner, F., Nösekabel, H., Kleinschmidt, P. (Hrsg.): *Multikonferenz Wirtschaftsinformatik 2006 (MKWI 06)*, Bd. 2, Berlin 2006, S. 33-43

LEWANDOWSKI 2008

Lewandowski, D.: *Suchmaschinenforschung im Kontext einer zukünftigen Webwissenschaft*. In: Scherfer, K. (Hrsg.): *Webwissenschaft - Eine Einführung*. Münster: LIT, 2008, S. 271

LEWANDOWSKI 2008B,

Lewandowski, D.: *The retrieval effectiveness of web search engines: considering results descriptions*. *Journal of Documentation* 64, 2008, 915-937.

LEWANDOWSKI 2009

Lewandowski, D.: *Standards der Ergebnispräsentation*. In: Lewandowski, D.: *Handbuch Internet Suchmaschinen*. Heidelberg: AKA Verlag, 2009, S.204

LEWANDOWSKI 2011

Lewandowski, D.: *Evaluierung von Suchmaschinen*. In: Lewandowski, D. (Hrsg.): *Handbuch Internet-Suchmaschinen 2: Neue Entwicklungen in der Web-Suche*. Heidelberg: Akademische Verlagsgesellschaft Aka, 2011, S. 206

## LEWANDOWSKI 2012

Lewandowski, D.: *A Framework for Evaluation the Retrieval Effectiveness of Search Engines*. In: Jouis, Christophe (Hrsg.): *Next Generation Search Engines: Advances Models for Information Retrieval*. Hershey, PA: IGI Global, 2012.

## LEWANDOWSKI 2013

Lewandowski, D.: *Suchmaschinen*. In: Kuhlen, R.; Semar, W.; Strauch, D.: *Grundlagen der praktischen Information und Dokumentation*. Berlin: de Gruyter, 2013. S. 496.

## LEWANDOWSKI/HÖCHSTÖTTER 2008

Lewandowski, D.; Höchstötter, N.: *Web Searching: A Quality Measurement Perspective*. In: Spink, A.; Zimmer, M. (Hrsg.): *Web Searching: Multidisciplinary Perspectives*. Berlin: Springer, 2008

## LORIGO et al. 2008

Lorigo, L.; Haridasan, M.; Brynjarsdottir, H.; Xia, L.; Joachims, T.; Gay, G.; Granka, L.; Pellacini, F. Pan, B.: *Eye tracking and online search: Lessons learned and challenges ahead*. *Journal of the American Society for Information Science and Technology* 59 (2008) 1041-1052, 2008

## NELSEN 2006

Nielsen, J.: *F-Shaped Pattern for Reading Web Content*. Alertbox [online] – URL: <http://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/> (Abruf: 2013-07-18)

## NICOLAI et al. 2009

Nicolai T.; Kirchhof, L.; Bruns, A.; Wilson, J.; Saunders, B.: *The self-Googleing phenomenon: Investigating the performance of personalized information resources*. 2009 In: *First Monday* [online] – URL:

<http://journals.uic.edu/ojs/index.php/fm/article/view/2683/2409> (Abruf: 2013-08-13)

PAN et al. 2007

Pan, B.; Hembrooke, H.; Joachims, T.; Lorigo, L.; Gay, G.; Granka, L.: *In Google we trust: Users' decisions on rank, position, and relevance*. Journal of Computer-Mediated Communication, 12(3), 2007

QUIRMBACH 2009

Lewandowski, D.: *Universal Search – Kontextuelle Einbindung von Ergebnissen unterschiedlicher Quellen und Auswirkungen auf das User Interface*. In: Lewandowski, D.: Handbuch Internet Suchmaschinen. Heidelberg: AKA Verlag, 2009, S.235

QUIRMBACH 2012

Quirmbach, S.: *Suchmaschinen: User Experience, Usability und nutzerzentrierte Website-Gestaltung*. Heidelberg: Springer, 2012, S. 131

RODDEN/FU 2006

Rodden, K.; Fu, X.: *Eyetracking Analysis of Google Web Search*. Internal Research, Google Inc. 2006

RODDEN/FU 2007

Rodden, K.; Fu, X.: *Exploring how mouse movements relate to eye movements on web search results pages*. Proceedings of the Conference on Research and Development. In: Information Retrieval, 2007

RÖßLER/UNGERER 2008

Rößler, I.; Ungerer, A.: *Statistik für Wirtschaftswissenschaftler*. Heidelberg, Springer 2008.

SARACEVIC 1999

Saracevic, T.: *Information science*. In: Journal of the American Society for Information Science 50, S. 1051-1063

STOCK 2007

Stock, W. G.: *Information Retrieval. Informationen suchen und finden*. Oldenbourg, München 2007, S. 68–81 (Kapitel 6, Relevanz und Pertinenz)

SÜNKLER 2012

Sünkler, S.: *Prototypische Entwicklung einer Software für die Erfassung und Analyse explorativer Suchen in Verbindung mit Tests zur Retrievaleffektivität*. Hamburg, 2012

TAGUE-SUTCLIFFE 1992

Tague-Sutcliffe, J.: *The Pragmatics of Information Retrieval Experimentation, revisited*. In: Information Processing & Management, 28(4), 1992, S. 467 – 490.

THUROW/MUSICA 2009

Thurrow, S.; Musica, N.: *When Search Meets Web Usability*, 1. Aufl.: New Riders, Berkeley 2009

TURPIN/SCHOLER 2006

Turpin, A.; Scholer, F.: *User Performance versus Precision Measures for Simple Search Tasks*. In: Proceedings of the 29th Annual International ACM Sigir Conference on Research and Development. Seattle, Washington, USA: Sigir '06, S. 11 – 18, 2006

USABILITY.DE 2009

Usability.de: *Goodbye Golden Triangle*. 2009. [online] – URL: <http://www.usability.de/publikationen/studien/goodbye-golden-triangle.html> (Abruf: 2013-08-13)

## VAN EIMEREN/FREES 2012

Van Eimeren, B.; Frees, B.: *Ergebnisse der ARD/ZDF-Onlinestudie 2012: 76% der Deutschen online – neue Nutzungssituationen durch mobile Endgeräte*. In: *Media Perspektiven* 7–8/2012, S. 369.

## VÉRONIS 2006

Véronis, J.: *A comparative study of six search engines*. 2006 [online] – URL: <http://www.up.univ-mrs.fr/veronis/pdf/2006-comparative-study.pdf> (Abruf: 2013-08-13)

## VOORHEES/HARMAN 2005

Voorhees, Ellen M.; Harman, Donna K.: *TREC : Experiment and Evaluation in Information Retrieval*. In: *Digital Libraries and Electronic Publishing*. London: The MIT Press, 2005

## WEBHITS.DE 2013

Web-Barometer [online]. – URL: <http://www.webhits.de/deutsch/webstats-.html/> (Abruf: 2013-07-18)

## WEB-STATS.INFO 2013

Suchmaschinen Marktanteile: Statistik für Juni 2013 [online]. – URL: <http://www.web-stats.info/blog/104-Suchmaschinen-Marktanteile-Statistik-fuer-Juni-2013.html> (Abruf: 2013-07-18)

**Anhang**

|  |    |
|--|----|
| Anhang 1: Fragebogen zur Ermittlung der Suchanfragen ..... | 86 |
| Anhang 2: Liste der verwendeten Suchanfragen .....         | 87 |
| Anhang 3: Daten und SPSS-Ergebnisse.....                   | 94 |

## Anhang 1: Fragebogen zur Ermittlung realer informationsorientierter Suchanfragen und den dazugehörigen Kontextinformationen

Meine letzte informationsorientierte Suchanfrage war:

---

... und das wollte ich wissen (1- bis 2 Sätze):

---

---

## Anhang 2: Liste der im Retrievaltest verwendeten Suchanfragen

| Suchanfrage                             | Hintergrund  |
|---|--|
| aldi angebote                           | Finde Informationen darüber, ob Aldi einen Online-Shop hat und welche Angebote es aktuell gibt.                                  |
| Allan Simonsen Unfall                   | Finde Informationen darüber, wie es zu den Unfall des dänischen Rennfahrer Allan Simonsen kommen konnte.                         |
| Altonale 2013                           | Finde Informationen wann und wo die Altonale dieses Jahr stattfindet.  |
| anfahrt Brandenburg                     | Finde Informationen, wie du am schnellsten nach Brandenburg mit dem Auto fahren kannst.  |
| angebrannter topf reinigen              | Finde Informationen darüber, wie sich ein angebrannter Kochtopf am besten reinigen lässt.  |
| Angela Merkel                           | Finde den vollen Namen von Angela Merkel.  |
| animal crossing new leaf                | Fine heraus, ob und wann das Konsolenspiel animal crossing im Handel erscheint.  |
| Anspruch auf Zwischenzeugnis            | Finde heraus, ob du als Angestellter einen rechtlichen Anspruch auf ein Zwischenzeugnis hast und wie du dabei vorgehen solltest. |
| apple tv 2 Anleitung                    | Finde eine Anleitung wie apple tv 2 richtig zu installieren ist.   |
| apploft Hamburg                         | Finde die Adresse der Firma apploft.   |
| Arbeitszeugnis Codes                    | Finde Informationen darüber, wie die von Personalern benutzen Codes in Arbeitszeugnissen zu interpretieren sind.                 |
| art 5 gg                                | Finde heraus, was art 5. GG besagt.  |
| audacity hintergrundgeräusche entfernen | Finde heraus wie man beim Audioeditor audacity Hintergrundgeräusche entfernt.  |
| axure import library                    | Finde heraus wie beim Prototyping-Programm axure eine Widget-Library importiert werden kann.                                     |
| bar definition                          | Finde eine Definition des Begriffes Bar (Gastronomie).   |
| Bartelstraße Hamburg                    | Finde heraus wo sich die Bartelstraße befindet und wie sie zu erreichen ist.   |
| Bedeutung Pfingsten                     | Finde Informationen zur Bedeutung von Pfingsten.   |
| Bedeutung Regenwald                     | Finde heraus welche Bedeutung der Regenwald für das Klima der Erde hat.  |
| benzinkosten Hamburg Nürnberg           | Finde Informationen darüber, wie hoch die Benzinkosten für die Strecke Hamburg-Nürnberg sind.                                    |
| Blocksatz                               | Finde heraus, was ein Blocksatz ist.   |
| blutspenden Hamburg                     | Finde Informationen darüber, wo man in Hamburg Blut spenden kann und wie man dabei vorgehen sollte.                              |
| Brasilien Sicherheit                    | Finde heraus, ob es sicher ist als Tourist nach Brasilien zu fliegen.  |
| Bundesliga Spielplan                    | Finde heraus, welche Vereine in der Bundesliga-Saison 2013/2014 spielen.   |
| Business Model Canvas                   | Finde heraus, was das Business Model Canvas ist und wie es aufgebaut ist.  |
| Campus Open Air Hamburg 2013            | Finde heraus wann das Campus Open Air Festival in Hamburg stattfindet und wie das Programm aussieht.                             |
| Chemiebranche Folgen Umwelt             | Finde Informationen darüber, welche Auswirkungen die Chemiebranche auf die Umwelt hat.   |
| Chris O'Dell                            | Finde heraus, wann der Autor Chris O'Dell lebte.   |
| clustering                              | Finde heraus, was clustering bedeutet.   |
| CMYK Einstellung PS6                    | Finde eine Anleitung, wie ein Bild vom Farbraum RGB auf CMYK bei Photoshop 6 umgewandelt wird.                                   |
| cookies löschen                         | Finde heraus wie Cookies aus dem Browserverlauf gelöscht werden.   |

|                                     |   |
|-------------------------------------|---|
| Cousine als Brautjungfer            | Finde Informationen darüber, ob man als Cousine der Braut gleichzeitig Brautjungfer sein darf.      |
| cpu                                 | Finde heraus was die Abkürzung CPU bedeutet.  |
| Damon Hill NRW Pedigree             | Finde das Pedigree einer Ahnentafel des Pferdes Damon Hill.   |
| DAX Börse                           | Finde den Börsenstand des DAX.  |
| Definition Palmöl                   | Finde heraus was Palmöl ist und wozu es verwendet wird.   |
| deutsche Gastarbeiter in polen      | Finde heraus ob es deutsche Gastarbeiter in Polen gibt und wo diese arbeiten.                       |
| distributionspolitik Bibliotheken   | Finde Informationen zu Distributionspolitik in Bibliotheken.  |
| diversity management Teilgebiete    | Finde heraus was die Teilgebiete von Diversity Management sind.                                     |
| dofollow                            | Finde heraus was der Begriff dofollow bedeutet.   |
| Dr. Nikolaus Töpfer                 | Finde die Adresse der HNO-Praxis von Dr. med. Nikolaus Töpfer.                                      |
| dreamweaver slideshow               | Finde Informationen darüber wie im Programm Dreamweaver eine Slideshow erstellt wird.               |
| dreamweaver video einbinden         | Finde Informationen darüber wie im Programm Dreamweaver ein Video eingebunden wird.                 |
| Durchlauferhitzer Mehrkosten        | Finde heraus ob ein Durchlauferhitzer Mehrkosten verursacht.  |
| Ebooksangebot Bibliotheken          | Finde heraus welche Bibliotheken eBooks haben und wie sie ausgeliehen werden können.                |
| Edward Snowden                      | Finde heraus wer Edward Snowden ist.  |
| Einkommensteuererklärung Muster     | Finde eine Vorlage/Muster für eine Einkommenssteuererklärung.                                       |
| Einwohnermeldeamt Hamburg           | Finde heraus wo sich das Einwohnermeldeamt in Hamburg befindet und wie die Öffnungszeiten lauten.   |
| Elbhochwasser Hamburg               | Finde heraus ob auch Hamburg von der Hochwasserkatastrophe betroffen sein könnte.                   |
| elbphilharmonie Hamburg Entwicklung | Finde heraus, wie sich der Bau der Elbphilharmonie in Hamburg weiterentwickelt hat.                 |
| erlebnisbad Ariba                   | Finde Informationen zur Anfahrt, Öffnungszeiten und Preise der Erlebnisbad Ariba.                   |
| error 1046 mysql datenbank          | Finde heraus, was der Fehler 1046 bei MySQL bedeutet und wie er behoben werden kann.                |
| ertragswertverfahren                | Finde Informationen darüber was ein Ertragswertverfahren ist.                                       |
| Eulers Identität                    | Finde heraus wie die Eulersche Formel lautet um eine Mathematikaufgabe zu lösen.                    |
| euro zloty                          | Finde heraus wie hoch der aktuelle Euro-Zloty Kurs liegt.   |
| excel makro erstellen               | Finde Informationen darüber wie in Excel Makros erstellt werden können.                             |
| facebook konto löschen              | Finde heraus ob ein Facebook-Account endgültig gelöscht werden kann und wie man dazu vorgehen muss. |
| Fahrgastfernsehen HVV               | Finde heraus wie das Fahrgastfernsehen des HVV produziert wird.                                     |
| faktor3 Adresse                     | Finde heraus wo sich der Sitz der PR-Agentur faktor3 befindet.                                      |
| Faszination sozialer Netzwerke      | Finde Informationen darüber, warum soziale Netzwerke unter Jugendlichen so beliebt sind.            |
| Fernsehprogramm                     | Finde das Fernsehprogramm um sich darüber zu informieren, welche Filme sehenswert sind.             |
| final cut pro rauschen entfernen    | Finde eine Anleitung wie Rauschen mit dem Videoschnittprogramm final cut pro entfernt wird.         |

|                              |   |
|------------------------------|---|
| final cut pro tutorial       | Finde ein Tutorial zur Nutzung des Videoschnittprogramms final cut pro.   |
| Fitzgerald                   | Finde heraus ob F. Scott Fitzgerald das Buch zu "Benjamin Button" geschrieben hat.  |
| gezi park                    | Finde Informationen darüber wie sich die Proteste in Istanbul um den Gezi Park entwickeln.  |
| giftige Lebensmittel Hund    | Finde Informationen darüber, welche Lebensmittel für Hunde giftig sind.   |
| Goldschatz von Eberswalde    | Finde Informationen über die Geschichte um den Goldschatz von Eberswalde.   |
| gti pleite                   | Finde heraus ob das Touristikunternehmen gti insolvent ist und welche Auswirkungen dies auf einen über das Unternehmen gebuchten Urlaub haben könnte. |
| GTI Travel                   | Finde heraus ob das Touristikunternehmen gti insolvent ist und welche Auswirkungen dies auf einen über das Unternehmen gebuchten Urlaub haben könnte. |
| Handball Damen Leipzig       | Finde heraus welche Damenhandballmannschaften es in Leipzig gibt.   |
| hangover 3 mr. chow          | Finde heraus wie der Schauspieler heißt, der im Film hangover 3 die Rolle des Mr. Chow spielt.  |
| harvard notation             | Finde heraus wie die Harvard Notation anzuwenden ist.   |
| hausmittel sonnenbrand       | Finde heraus welche Hausmittel gegen Sonnenbrand helfen sollen.   |
| HAW Prüfungstermine          | Finde heraus wann an der HAW genau die Prüfungstermine sind.  |
| HAW Semesterzeiten           | Finde heraus wann an der HAW die vorlesungsfreie Zeit beginnt.  |
| hjerteknuser translation     | Finde heraus worüber die norwegische Band Kaizers Orchestra im Song hjerteknuser singt.   |
| hochwasser hamburg           | Finde heraus ob Hamburg durch das Hochwasser gefährdet ist.   |
| Hochwasser Lüneburg          | Finde heraus ob Lüneburg durch das Hochwasser betroffen ist.  |
| Hochwasser Wittenberge       | Finde heraus wie die aktuelle Hochwassersituation in der Stadt Wittenberge ist.   |
| hostel berlin                | Finde heraus wo welche Hotels in Berlin zu finden sind und wie viel sie kosten.   |
| hundenauslaufzonen hamburg   | Finde heraus ob es in Hamburg Hundenauslaufzonen gibt und wie sie zu erreichen sind.  |
| Hurricane Festival 2013      | Finde Informationen wann das Hurricane Festival dieses Jahr stattfindet.  |
| hurricane plan               | Finde einen Plan über die Veranstaltungen beim Hurricane Festival.  |
| indikativ                    | Finde heraus was das Adjektiv indikativ bedeutet.   |
| Ingredienzien                | Finde heraus wie man das Wort Ingredienzien richtig schreibt.   |
| integrierte Kommunikation    | Finde eine Definition des Begriffes integrierte Kommunikation.  |
| ipad hülle tastatur          | Finde heraus ob es für das iPad eine Hülle mit integrierter Tastatur gibt.  |
| iron man 3 kritik            | Finde heraus ob der Film Iron Man 3 sehenswert ist.   |
| Jeanete Walls Roman          | Finde heraus wann das Fortsetzungswerk der Autorin Jeanette Walls in den Handel kommt.  |
| Jordanische Botschaft Berlin | Finde heraus wo sich die jordanische Botschaft in Hamburg befindet.   |
| kanye kim baby               | Finde heraus wie das Kind des Musikers Kanye West   |

---

|                                      |  |
|--------------------------------------|--|
|                                      | und des IT-Girls Kim Kardashian heißt.   |
| Kieler Woche                         | Finde heraus wann die Kieler Woche stattfindet und wie das Programm aussieht.                  |
| klettern hamburg                     | Finde heraus ob es in Hamburg Kletterparks gibt.   |
| Kontakt Google Hamburg               | Finde heraus wie du mit Google Hamburg in Kontakt treten kannst.                               |
| Kriterien Wikipediaartikel           | Finde heraus welche Kriterien ein Wikipediaartikel erfüllen muss, um veröffentlicht zu werden. |
| Kulturufer Friedrichshafen           | Finde heraus welche Veranstaltungen und Bands beim diesjährigen Kulturufer dabei sind.         |
| leserverhalten                       | Finde heraus welche Studien es zum Thema Leserverhalten gibt.                                  |
| library 2.0                          | Finde Webseiten zum Thema Kataloganreicherung und Web 2.0 Anwendungen in Bibliothek-Katalogen. |
| literarischer Journalismus           | Finde Informationen zum Thema literarischer Journalismus für eine Hausarbeit.                  |
| lyrics radioactive                   | Finde heraus wie der Songtext der Single Radioactive von Imagine Dragons lautet.               |
| Mai Tai rezept                       | Finde ein Rezept um den Cocktail Mai Tai zu machen.  |
| man of steel clark kent              | Finde heraus wie der Schauspieler heißt der im Film Man of Steel die Rolle Clark Kent spielt.  |
| master hamburg                       | Finde heraus welche Masterstudiengänge es an hamburger Unis/Hochschulen gibt.                  |
| masterstudiengänge hamburg           | Finde heraus welche Masterstudiengänge es an hamburger Unis/Hochschulen gibt.                  |
| medien praktikum hamburg             | Finde heraus wo in Hamburg ein Praktikum in einen Medienunternehmen gemacht werden kann.       |
| medienkompetenz                      | Finde eine Definition für den Begriff Medienkompetenz.   |
| medienservice finkenau               | Finde heraus wann der Medienservice Finkenau geöffnet hat.                                     |
| melone feta salat                    | Finde ein Rezept für einen Salat aus Fetakäse und Melone.                                      |
| ministerpräsident schleswig holstein | Finde heraus wie der Ministerpräsident von Schleswig-Holstein heißt.                           |
| minze abwaschen                      | Finde heraus, ob Minze vor dem Verzehr abgewaschen werden muss.                                |
| nelson mandela                       | Finde Informationen über den Gesundheitszustand von Nelson Mandela                             |
| nihilismus                           | Finde eine Definition für den Begriff Nihilismus   |
| nutzwertanalyse beispiele            | Finde Beispiele wie eine Nutzerwertanalyse durchzuführen ist.                                  |
| Obama rede berlin                    | Finde heraus was Barack Obama in seiner Rede in Berlin sagte.                                  |
| Oliver Kreuzer Karriere              | Finde heraus was für Stationen Oliver Kreuzer in seiner Karriere als Sportmanager hatte.       |
| online marketing hamburg             | Finde heraus welche Online Marketing Agenturen es in Hamburg gibt.                             |
| online marketing manager Berufsbild  | Finde heraus was Online Marketing Manager für eine Tätigkeit haben.                            |
| ordentliche Kündigung                | Finde heraus was eine ordentliche Kündigung ist und wann sie rechtskräftig ist.                |
| pdf zusammenfügen                    | Finde heraus wie du mehrere PDF Dokumente zu einen zusammenfügen kannst.                       |
| pflege miniorchideen                 | Finde heraus wie Mini-Orchideen gepflegt werden müssen.  |
| pokemon schwarz cheats               | Finde ein cheat bzw. eine Anleitung um das Computerspiel Pokemon Schwarz zu beenden.           |

---

|                                      |  |
|--------------------------------------|--|
| ppt datei wiederherstellen           | Finde heraus ob und wie eine Powerpointdatei wiederhergestellt werden kann, nachdem das Programm abgestürzt ist. |
| programm cinemax dammtor             | Finde heraus welche Filme im Kino Cinemaxx am Dammtor laufen.  |
| PS4                                  | Finde Informationen zur Spielekonsole Playstation 4  |
| ps4 xbox one                         | Finde einen Vergleich zwischen den Spielekonsolen Playstation 4 und X Box one                                    |
| Pyramidenform Journalismus           | Finde heraus welche Bestandteile die Journalismus-Pyramide hat.  |
| Qualitätslenkung                     | Finde Informationen zum Thema Qualitätslenkung für ein Referat.  |
| raffaello werbung song 2013          | Finde heraus wie das Lied aus der aktuellen Raffaello Werbung heißt.   |
| raw datei                            | Finde heraus was eine RAW-Datei ist.   |
| Recommender Systeme                  | Finde heraus was Recommendr Systeme sind.  |
| reiseversicherungsschein             | Finde heraus was ein Reiseversicherungsschein ist.   |
| retrievalmodelle                     | Finde Informationen zum Thema Retrievalmodelle   |
| retro strahlen effekt photospho      | Finde heraus wie man mit dem Programm Photoshop den Effekt retro Strahlen erzeugen kann.                         |
| rezept couscous                      | Finde ein Rezeot für Couscous  |
| Rezept Creme Brulee                  | Finde heraus welche Zutaten zum Zubereiten einer Creme Brulee nötig sind.  |
| rezept erdbeerkuchen                 | Finde ein Rezept für Erdbeerkuchen   |
| rezept quiche lorraine               | Finde ein Rezept für Quicke Lorraine   |
| Rezept Spargel                       | Finde ein Rezept für Spargelgerichte   |
| robert lewandowski wechsel           | Finde Informationen darüber ob Robert Lewandowski zum FC Bayern wechseln wird                                    |
| Robin Hood                           | Finde Informationen zur Legende von Robin Hood.  |
| router überspannungsschaden gewitter | Finde heraus ob bei einem kaputten Router durch Überspannung die Versicherung die Kosten übernimmt.              |
| samsung galaxy s4 erfahrungen        | Finde heraus ob das Smartphone Samsung Galaxy s4 gut bewertet wird.  |
| sara paxton                          | Finde heraus wann Sara Paxton geboren wurde und welchen Beruf sie hat  |
| seelen kinostart                     | Finde heraus wann der Film Seelen in die Kinos kommt.  |
| sehenswürdigkeiten kopenhagen        | Finde heraus welche Sehenswürdigkeiten Kopenhagen hat und wie sie zu erreichen sind.                             |
| sehenswürdigkeiten kreta             | Finde heraus welche Sehenswürdigkeiten Kreta hat und wie sie zu erreichen sind.                                  |
| shitstorm                            | Finde eine Definition zum Begriff Shitstorm  |
| Snowden                              | Finde heraus wer Edward Snowden ist.   |
| social media marketing               | Finde Informationen zum Thema Social Media Marketing für eine Hausarbeit   |
| social networking                    | Finde eine Definition für den Begriff Social Networking  |
| soleil moon frye height              | Finde heraus wie groß die Schauspielerin Soleil Moon-Frye ist.   |
| Sommerferien Hamburg                 | Finde heraus wann der letzte Schultag in Hamburg ist.  |
| Sommerferien Niedersachsen           | Finde heraus wann die Sommerferien 2013 in Niedersachsen beginnen.   |
| song planet sports werbung           | Finde heraus wie der Song aus der aktuellen Planet Sports Fernsehwerbung heißt                                   |
| Sonnenstudio Hamburg                 | Finde heraus wo sich in Hamburg Sonnenstudios befinden.  |
| sopio                                | Finde heraus was sopio ist.  |
| spiegel bürgerschaft ham-            | Finde heraus ob es einen Spiegel-Artikel von der Bür-  |

|   |  |
|---|--|
| burg  | gerschaftssitzung vom 19.06.13 gibt.   |
| Stadtfest Ahrensburg Programm                 | Finde heraus was auf dem Stadtfest in Ahrensburg geboten wird.   |
| stiftstraße 69                                | Finde heraus wo sich die Stiftstraße in Hamburg befindet.  |
| student steuererklärung                       | Finde heraus ob man als Student eine Steuererklärung machen muss.                                      |
| studentenjobs hamburg                         | Finde ein Portal für Studentenjobs in Hamburg.   |
| suchmaschinenoptimierung hamburg              | Finde heraus welche SEO-Agenturen es in Hamburg gibt.  |
| Suits Episode 3 USA                           | Finde heraus wann die 3 Staffel der Serie Suits in den USA startet.                                    |
| sushi hamburg                                 | Finde heraus wo man in Hamburg sushi essen kann.   |
| sylt bam-bus klaus                            | Finde heraus ob es die Baum bus Bar auf Sylt noch nach dem Tod des Besitzers gibt                      |
| synonym auswertbar                            | Finde heraus welche Synonyme es für das wort auswertbar gibt.  |
| synonym ressourcen                            | Finde heraus welche Synonyme es für das wort Ressourcen gibt.  |
| SysVinit                                      | Finde eine Definition für das Wort SysVinit  |
| talstraße hamburg                             | Finde heraus wo sich die talstraße in Hamburg befindet.  |
| The Great Gatsby                              | Finde Informationen darüber, ob der Film The Great Gatsby auf einem Buch basiert.                      |
| The Last of Us                                | Finde Informationen und Erfahrungen zum Computerspiel The Last of Us                                   |
| the voice uk 2013 ep 4                        | Finde heraus wie die Teilnehmer und Juroren aus der Britischen Talentshow The Voice (Episode 4) heißen |
| thesaurus                                     | Finde heraus was ein Thesaurus ist   |
| tk garnelen zubereiten                        | Finde heraus wie tiefgefrorene Garnelen zubereitet werden müssen.                                      |
| top level domain                              | Finde heraus was eine Top Level Domain ist.  |
| Untermieterzuschlag                           | Finde heraus wann ein Untermieterzuschlag rechtskräftig ist.   |
| unterschied thesaurus klassifikation          | Finde heraus was der Unterschied zwischen Thesaurus und Klassifikation ist.                            |
| urlaub 400 euro job                           | Finde heraus ob man als geringfügig Beschäftigter Anspruch auf Urlaub hat und wie lange.               |
| urlaubsziele november                         | Finde heraus welche Urlaubsländer im November gutes (warmes) Wetter bieten.                            |
| usability agentur hamburg                     | Finde heraus ob und welche Usability-Agenturen es in Hamburg gibt.                                     |
| user centred design                           | Finde eine Definition für User Centred Design.   |
| vapiano                                       | Finde heraus wo sich die Filialen der Restaurant-Kette Vapiano befinden und wann sie offen haben.      |
| vegan restaurant hamburg                      | Finde heraus wo man in Hamburg Vegan essen gehen kann.   |
| vegetarisch grillen                           | Finde Rezepte für vegetarische Grillspeisen  |
| vergleichsstudie informationsverhalten medien | Finde heraus ob es Vergleichsstudien zum Thema Informationsverhalten in den Medien gibt                |
| verkaufsoffener Sonntag hamburg               | Finde heraus wann der nächste verkaufsoffene Sonntag in Hamburg stattfindet                            |
| Victor Murdock                                | Finde Steckbriefinformationen über den US Politiker Victor Murdock                                     |
| Vogelnest ohne Eltern                         | Finde heraus was zu tun ist, wenn ein Vogelnest ohne die Elterntiere gefunden worden ist               |
| vollmond                                      | Finde heraus wann Vollmond ist   |
| volltextindexierung                           | Finde heraus wie Volltextindexierung funktioniert  |

---

|                                     |   |
|-------------------------------------|---|
| Walter Jens Nachruf                 | Finde Zeitungsartikel zum Nachruf von Walter Jens   |
| Wandern Sarek                       | Finde Reiseberichte Tipps zum Wandern/Trekking im Sarek Nationalpark (Schweden)             |
| was hilft gegen muskelkater         | Finde Tipps was gegen Muskelkater wirkt   |
| Web 2.0 Anwendungen in Bibliotheken | Finde heraus welche Web 2.0 Anwendungen es in Bibliotheken gibt.                            |
| werkstudent krankversicherung       | Finde heraus ob man als Werkstudent krankversicherungspflichtig ist.                        |
| werkstudent rentenversicherung      | Finde heraus ob man als Werkstudent rentenversicherungspflichtig ist.                       |
| wetter amsterdam                    | Informiere dich über das Wetter in Amsterdam  |
| wetter hamburg                      | Informiere dich über das Wetter in Hamburg  |
| wetter soltau                       | Informiere dich über das Wetter in Soltau   |
| Wii U                               | Finde Informationen zur Spielkonsole Wii U und ob sie empfohlen wird.                       |
| wohnung untervermieten              | Finde heraus ob man eine Mietwohnung untermieten darf und was zu beachten ist               |
| word silbentrennung                 | Finde heraus wo in Word eine automatische Silbentrennung eingestellt wird.                  |
| wordpress auf server installieren   | Finde heraus wie Wordpress als Content Management System auf einen Server installiert wird. |
| wordpress blog backup               | Finde heraus wie man für einen Wordpress Blog ein Backup erstellt.                          |
| zitat musikindustrie eigentum       | Finde das Zitat von Reginald Rudolf zur Musikindustrie                                      |
| zoll in cm                          | Finde eine Umrechnung von Zoll in cm  |
| zubereitung spargel                 | Finde heraus wie Spargel zubereitet wird  |

---

## Anhang 3: Daten und SPSS-Ergebnisse

**Häufigkeiten für alle Suchanfragen und Positionen**

|        |    | Häufigkeit | Prozent | Gültige Prozent | Kumulative Prozente |
|--------|----|------------|---------|-----------------|---------------------|
| Gültig | 1  | 804        | 11,2    | 11,7            | 11,7                |
|        | 2  | 92         | 1,3     | 1,3             | 13,1                |
|        | 3  | 223        | 3,1     | 3,3             | 16,3                |
|        | 4  | 154        | 2,1     | 2,2             | 18,6                |
|        | 5  | 167        | 2,3     | 2,4             | 21,0                |
|        | 6  | 177        | 2,5     | 2,6             | 23,6                |
|        | 7  | 152        | 2,1     | 2,2             | 25,8                |
|        | 8  | 143        | 2,0     | 2,1             | 27,9                |
|        | 9  | 137        | 1,9     | 2,0             | 29,9                |
|        | 10 | 116        | 1,6     | 1,7             | 31,6                |
|        | 11 | 91         | 1,3     | 1,3             | 32,9                |
|        | 12 | 103        | 1,4     | 1,5             | 34,4                |
|        | 13 | 101        | 1,4     | 1,5             | 35,9                |
|        | 14 | 102        | 1,4     | 1,5             | 37,3                |
|        | 15 | 72         | 1,0     | 1,0             | 38,4                |
|        | 16 | 54         | ,8      | ,8              | 39,2                |
|        | 17 | 64         | ,9      | ,9              | 40,1                |
|        | 18 | 56         | ,8      | ,8              | 40,9                |
|        | 19 | 54         | ,8      | ,8              | 41,7                |
|        | 20 | 61         | ,8      | ,9              | 42,6                |
|        | 21 | 51         | ,7      | ,7              | 43,4                |
|        | 22 | 49         | ,7      | ,7              | 44,1                |
|        | 23 | 47         | ,7      | ,7              | 44,8                |
|        | 24 | 59         | ,8      | ,9              | 45,6                |
|        | 25 | 50         | ,7      | ,7              | 46,3                |
|        | 26 | 49         | ,7      | ,7              | 47,1                |
|        | 27 | 40         | ,6      | ,6              | 47,6                |
|        | 28 | 34         | ,5      | ,5              | 48,1                |
|        | 29 | 38         | ,5      | ,6              | 48,7                |
|        | 30 | 43         | ,6      | ,6              | 49,3                |
|        | 31 | 47         | ,7      | ,7              | 50,0                |
|        | 32 | 48         | ,7      | ,7              | 50,7                |
|        | 33 | 44         | ,6      | ,6              | 51,3                |
|        | 34 | 36         | ,5      | ,5              | 51,9                |
|        | 35 | 37         | ,5      | ,5              | 52,4                |
|        | 36 | 33         | ,5      | ,5              | 52,9                |
|        | 37 | 44         | ,6      | ,6              | 53,5                |
|        | 38 | 37         | ,5      | ,5              | 54,1                |
|        | 39 | 42         | ,6      | ,6              | 54,7                |

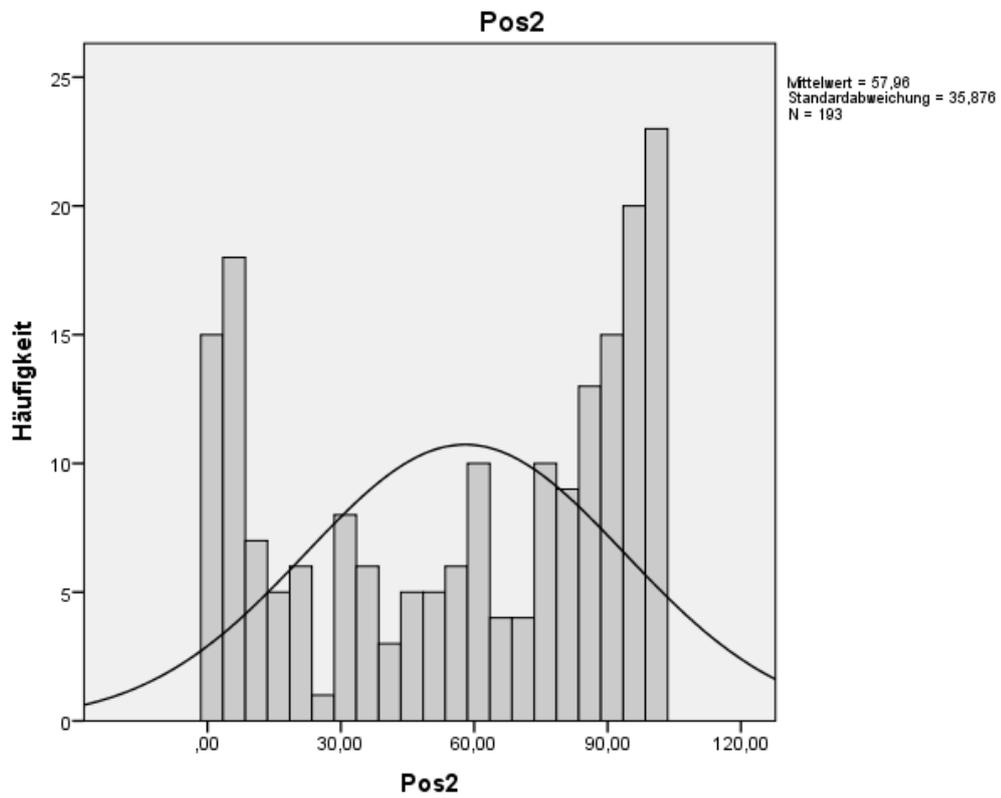
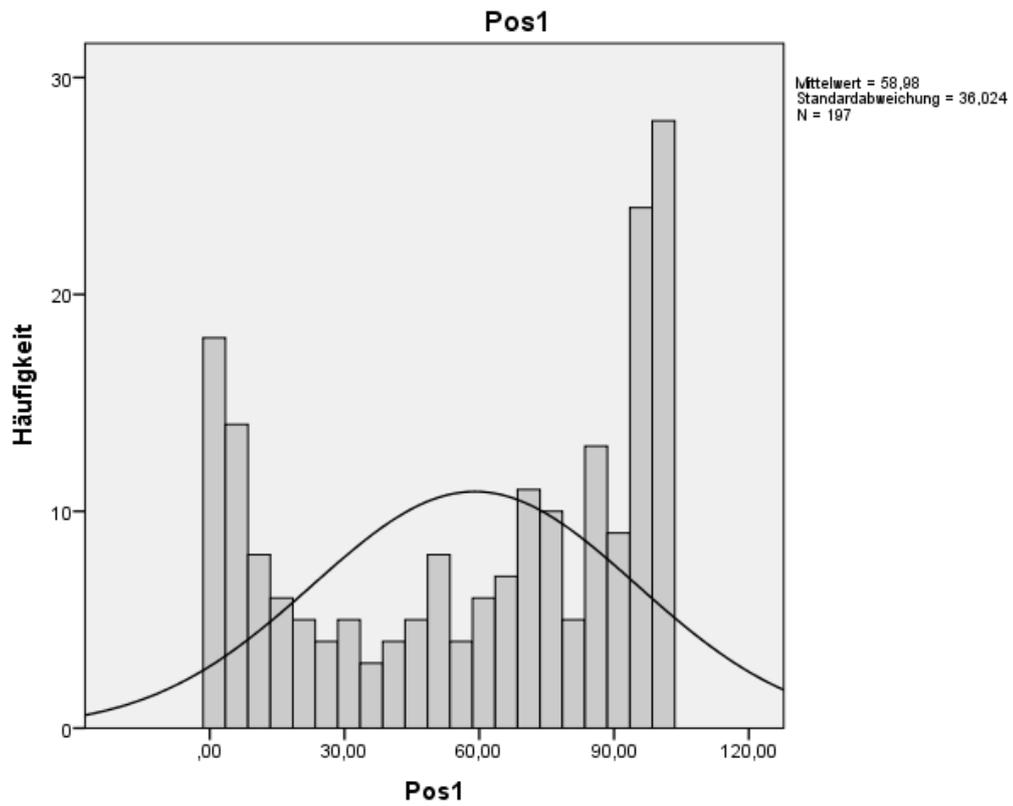
---

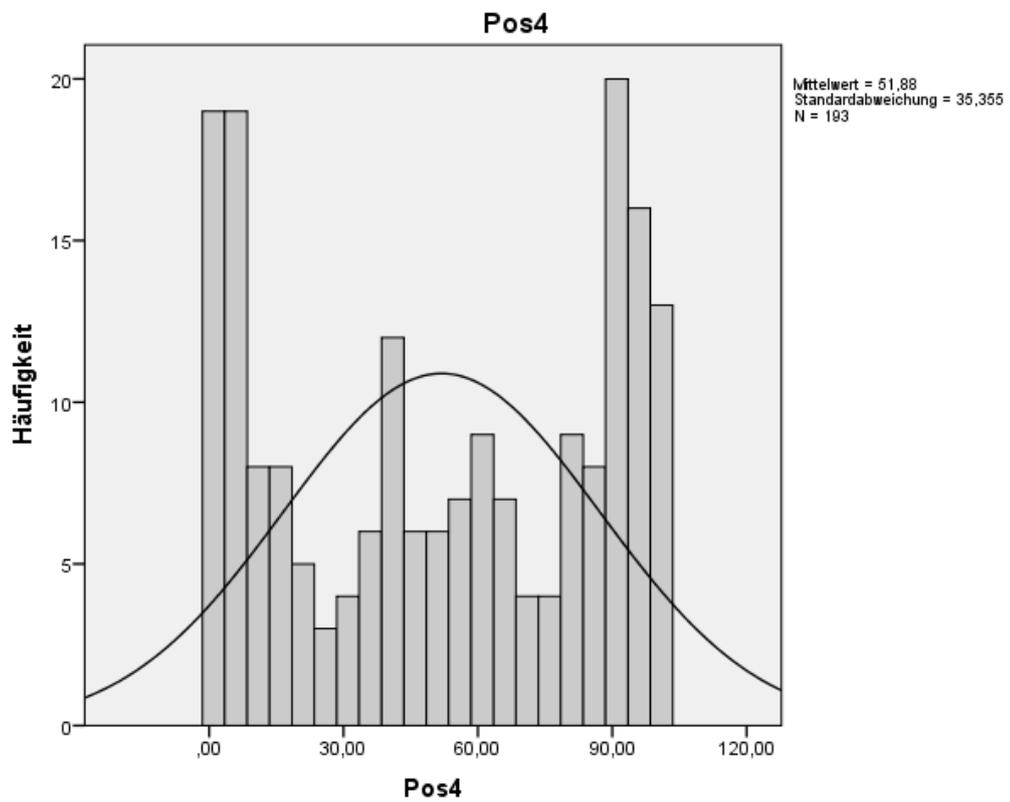
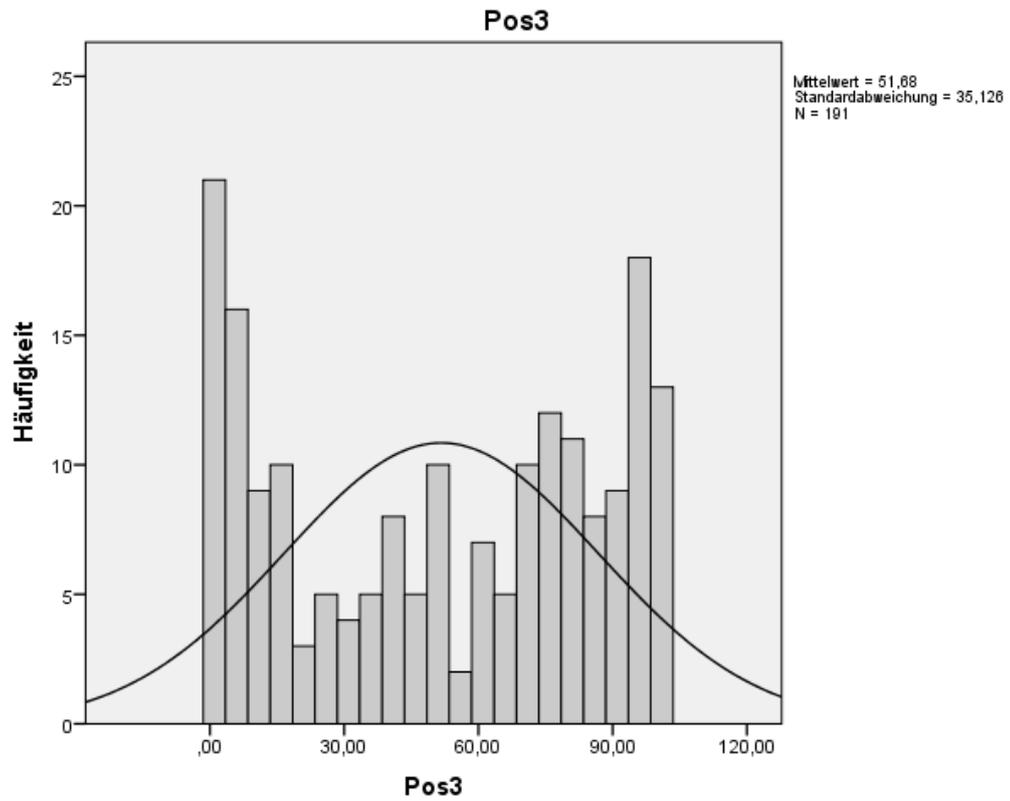
|    |    |    |    |      |
|----|----|----|----|------|
| 40 | 38 | ,5 | ,6 | 55,2 |
| 41 | 58 | ,8 | ,8 | 56,1 |
| 42 | 36 | ,5 | ,5 | 56,6 |
| 43 | 43 | ,6 | ,6 | 57,2 |
| 44 | 48 | ,7 | ,7 | 57,9 |
| 45 | 29 | ,4 | ,4 | 58,4 |
| 46 | 35 | ,5 | ,5 | 58,9 |
| 47 | 43 | ,6 | ,6 | 59,5 |
| 48 | 44 | ,6 | ,6 | 60,1 |
| 49 | 52 | ,7 | ,8 | 60,9 |
| 50 | 64 | ,9 | ,9 | 61,8 |
| 51 | 44 | ,6 | ,6 | 62,5 |
| 52 | 58 | ,8 | ,8 | 63,3 |
| 53 | 48 | ,7 | ,7 | 64,0 |
| 54 | 35 | ,5 | ,5 | 64,5 |
| 55 | 51 | ,7 | ,7 | 65,3 |
| 56 | 36 | ,5 | ,5 | 65,8 |
| 57 | 40 | ,6 | ,6 | 66,4 |
| 58 | 48 | ,7 | ,7 | 67,1 |
| 59 | 54 | ,8 | ,8 | 67,9 |
| 60 | 52 | ,7 | ,8 | 68,6 |
| 61 | 56 | ,8 | ,8 | 69,4 |
| 62 | 50 | ,7 | ,7 | 70,2 |
| 63 | 46 | ,6 | ,7 | 70,8 |
| 64 | 49 | ,7 | ,7 | 71,5 |
| 65 | 56 | ,8 | ,8 | 72,4 |
| 66 | 51 | ,7 | ,7 | 73,1 |
| 67 | 42 | ,6 | ,6 | 73,7 |
| 68 | 36 | ,5 | ,5 | 74,2 |
| 69 | 54 | ,8 | ,8 | 75,0 |
| 70 | 46 | ,6 | ,7 | 75,7 |
| 71 | 44 | ,6 | ,6 | 76,3 |
| 72 | 45 | ,6 | ,7 | 77,0 |
| 73 | 48 | ,7 | ,7 | 77,7 |
| 74 | 48 | ,7 | ,7 | 78,4 |
| 75 | 54 | ,8 | ,8 | 79,2 |
| 76 | 34 | ,5 | ,5 | 79,7 |
| 77 | 43 | ,6 | ,6 | 80,3 |
| 78 | 34 | ,5 | ,5 | 80,8 |
| 79 | 43 | ,6 | ,6 | 81,4 |
| 80 | 40 | ,6 | ,6 | 82,0 |
| 81 | 40 | ,6 | ,6 | 82,6 |
| 82 | 38 | ,5 | ,6 | 83,1 |
| 83 | 45 | ,6 | ,7 | 83,8 |

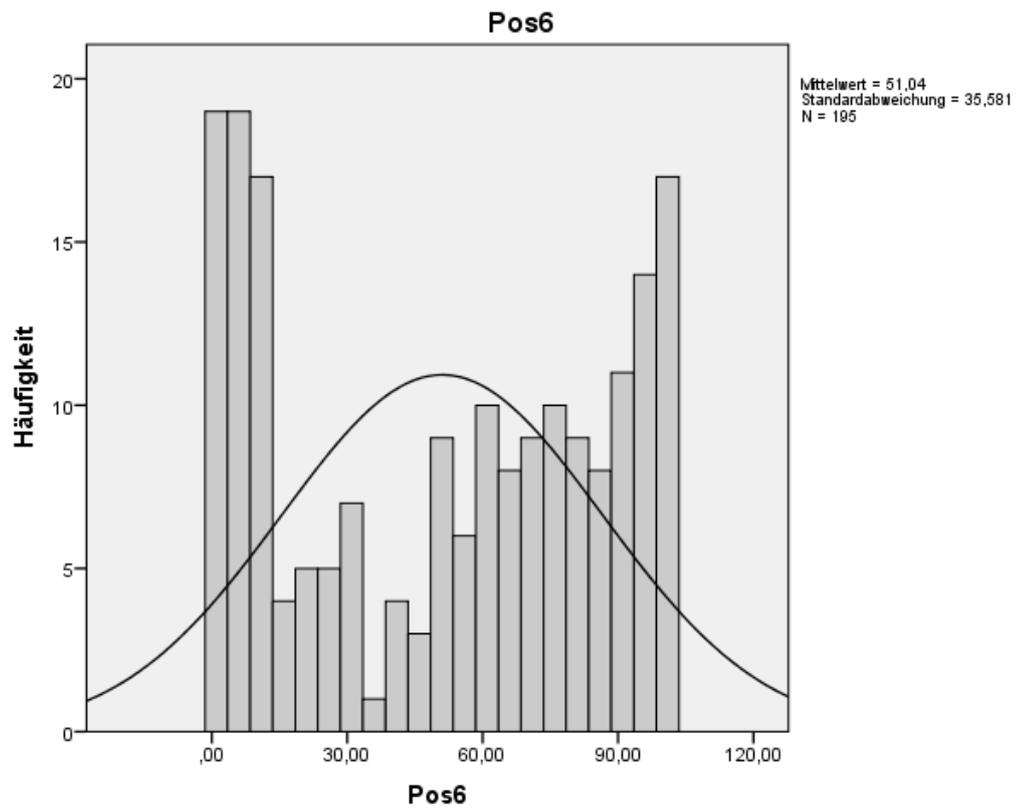
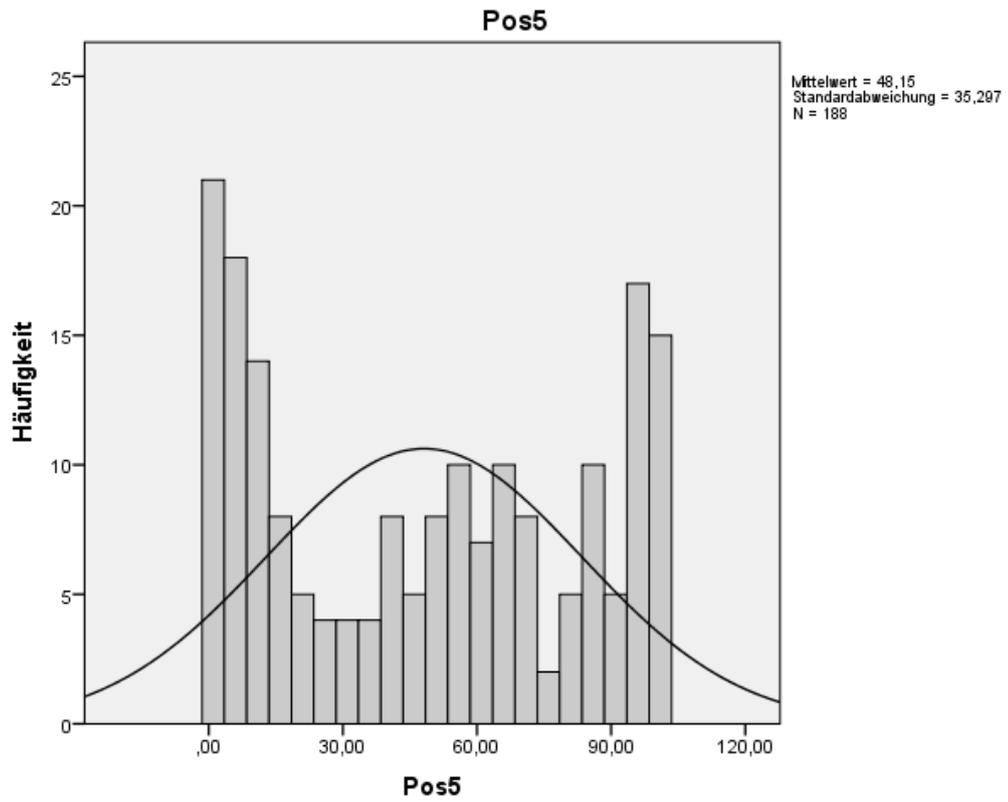
|             |      |       |       |       |
|-------------|------|-------|-------|-------|
| 84          | 41   | ,6    | ,6    | 84,4  |
| 85          | 37   | ,5    | ,5    | 84,9  |
| 86          | 41   | ,6    | ,6    | 85,5  |
| 87          | 49   | ,7    | ,7    | 86,3  |
| 88          | 51   | ,7    | ,7    | 87,0  |
| 89          | 50   | ,7    | ,7    | 87,7  |
| 90          | 49   | ,7    | ,7    | 88,4  |
| 91          | 57   | ,8    | ,8    | 89,3  |
| 92          | 48   | ,7    | ,7    | 90,0  |
| 93          | 58   | ,8    | ,8    | 90,8  |
| 94          | 48   | ,7    | ,7    | 91,5  |
| 95          | 62   | ,9    | ,9    | 92,4  |
| 96          | 48   | ,7    | ,7    | 93,1  |
| 97          | 59   | ,8    | ,9    | 94,0  |
| 98          | 88   | 1,2   | 1,3   | 95,3  |
| 99          | 103  | 1,4   | 1,5   | 96,8  |
| 100         | 222  | 3,1   | 3,2   | 100,0 |
| Gesamtsumme | 6860 | 95,5  | 100,0 |       |
| Fehlend 0   | 322  | 4,5   |       |       |
| Gesamtsumme | 7182 | 100,0 |       |       |

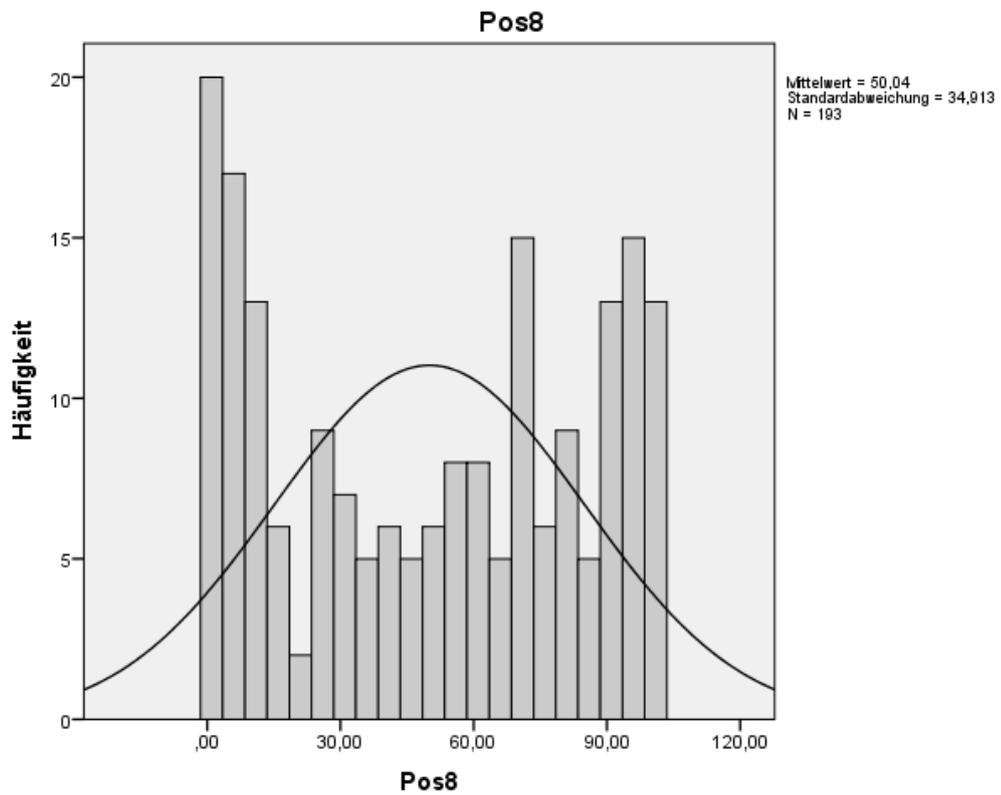
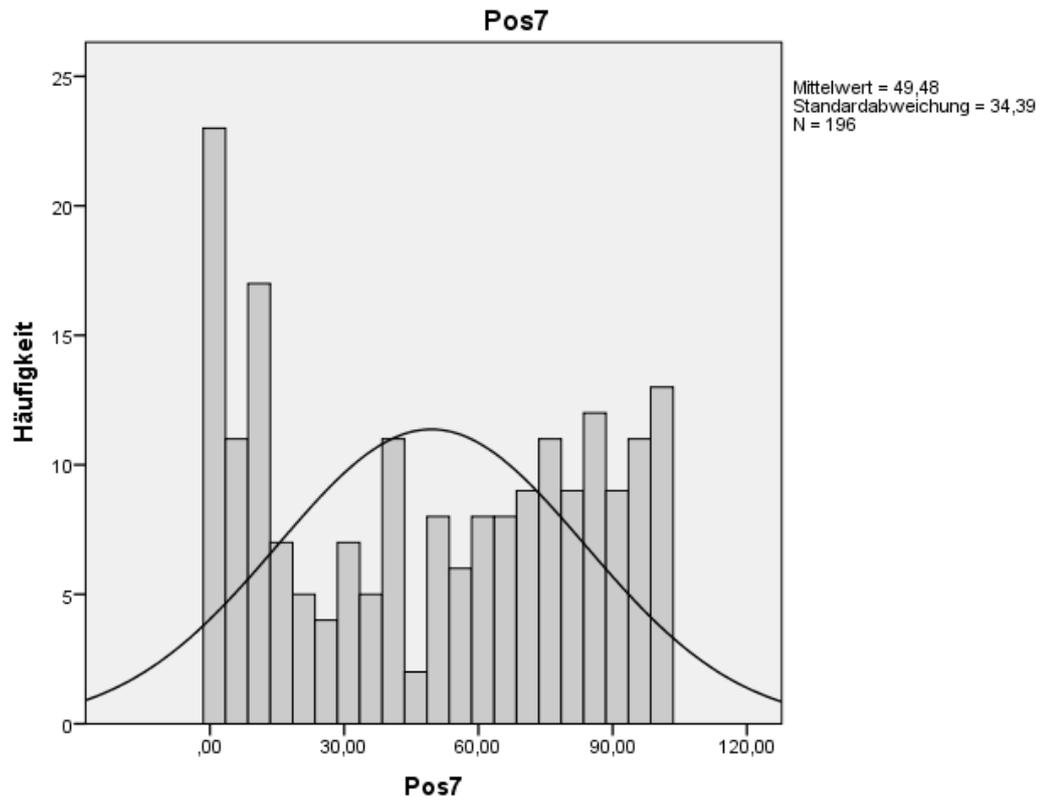
### Deskriptive Statistik

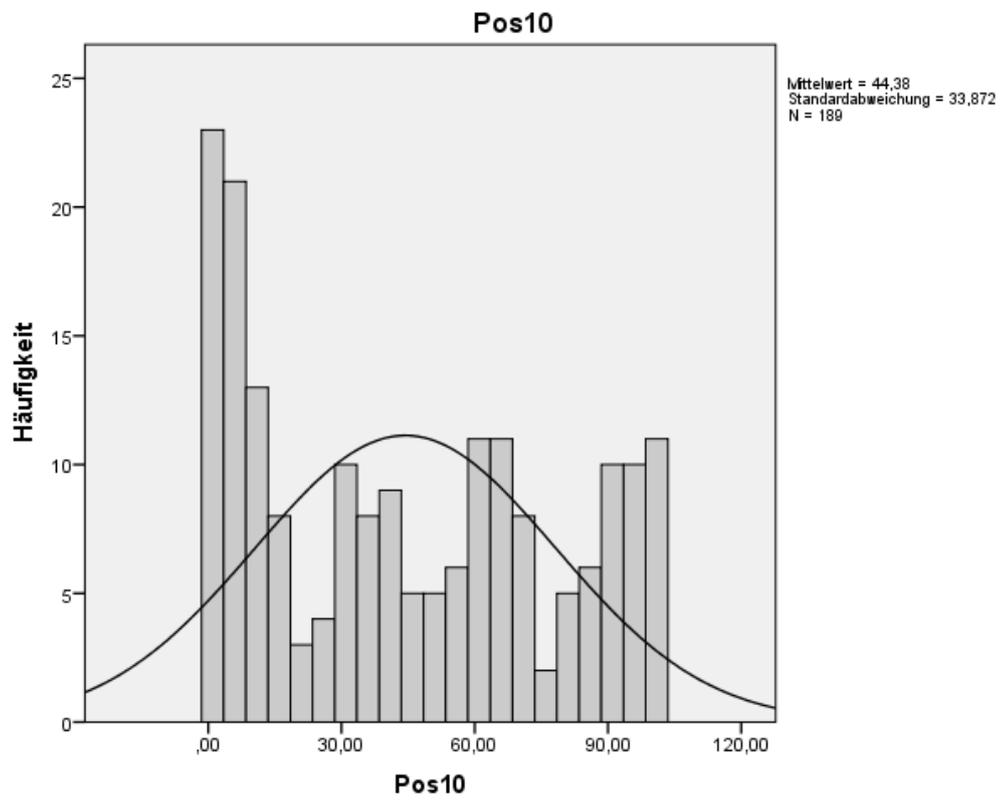
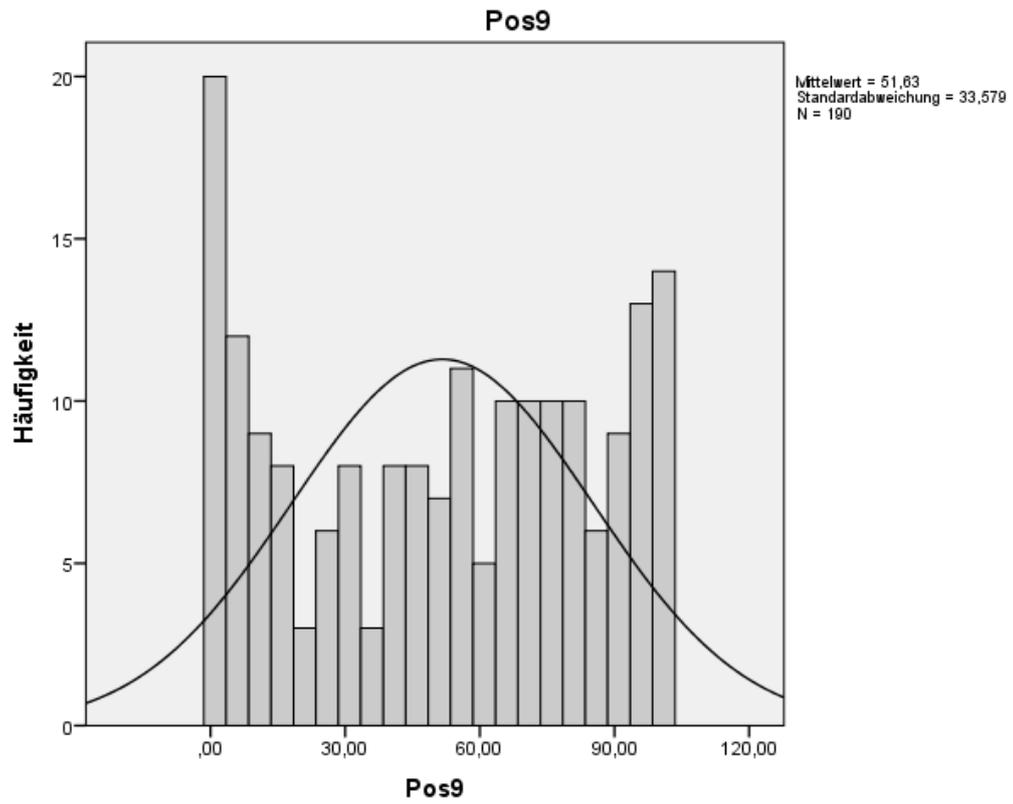
|                             |         |          |
|-----------------------------|---------|----------|
| N                           | Gültig  | 6860     |
|                             | Fehlend | 322      |
| Mittelwert                  |         | 39,57    |
| Median                      |         | 31,50    |
| Modalwert                   |         | 1        |
| Standardabweichung          |         | 33,950   |
| Varianz                     |         | 1152,596 |
| Schiefe                     |         | ,418     |
| Standardfehler der Schiefe  |         | ,030     |
| Kurtosis                    |         | -1,286   |
| Standardfehler der Kurtosis |         | ,059     |
| Summe                       |         | 271451   |

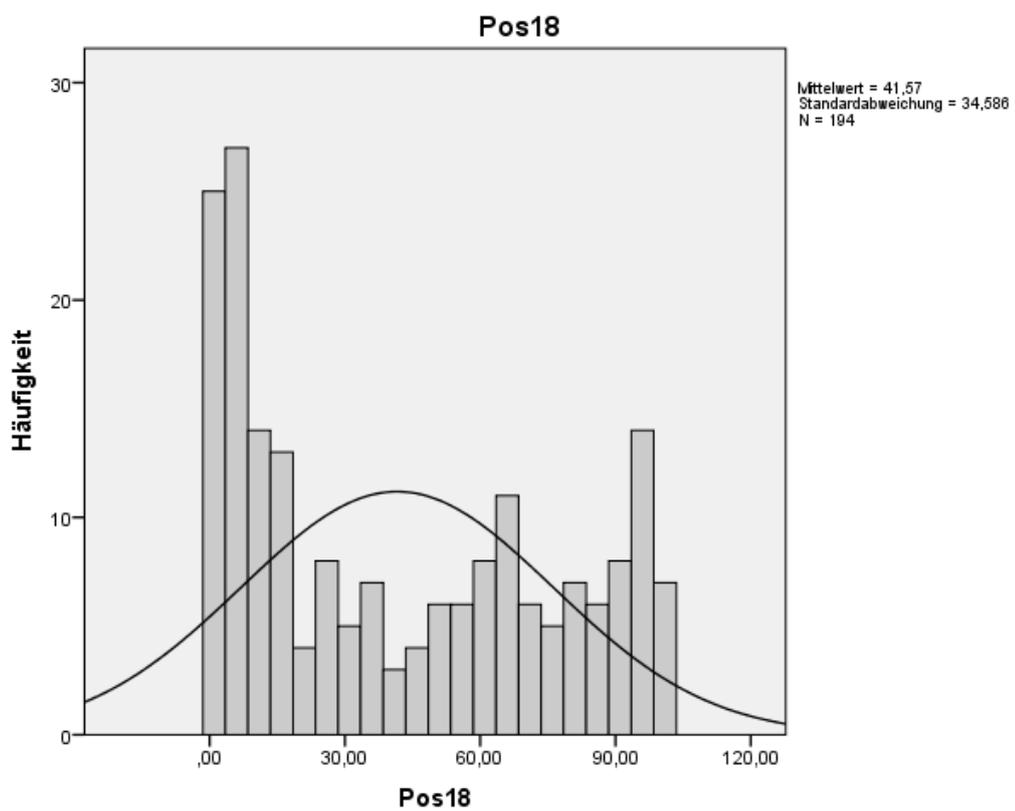
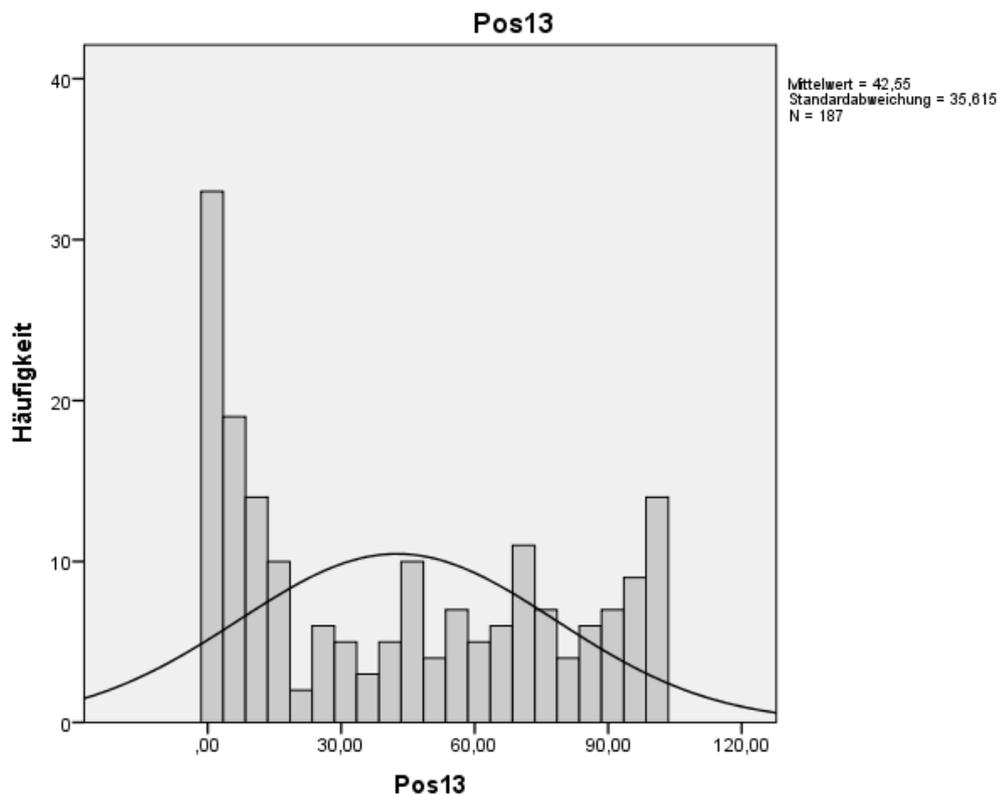
**Histogramme Häufigkeiten pro Position**

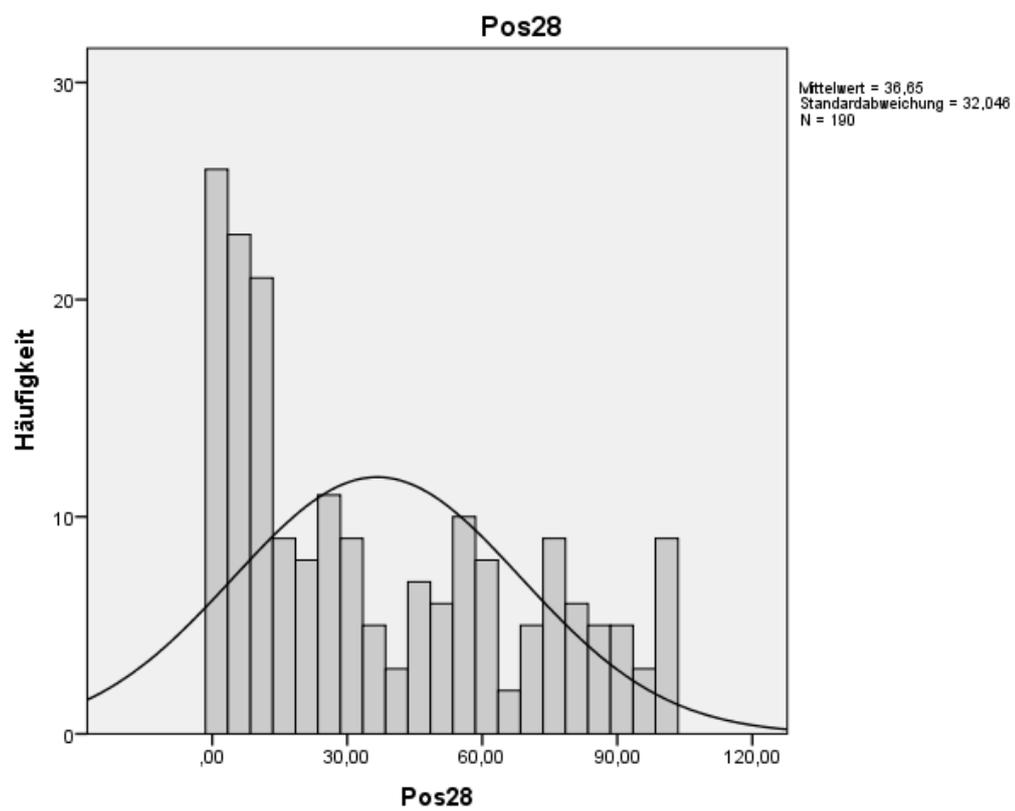
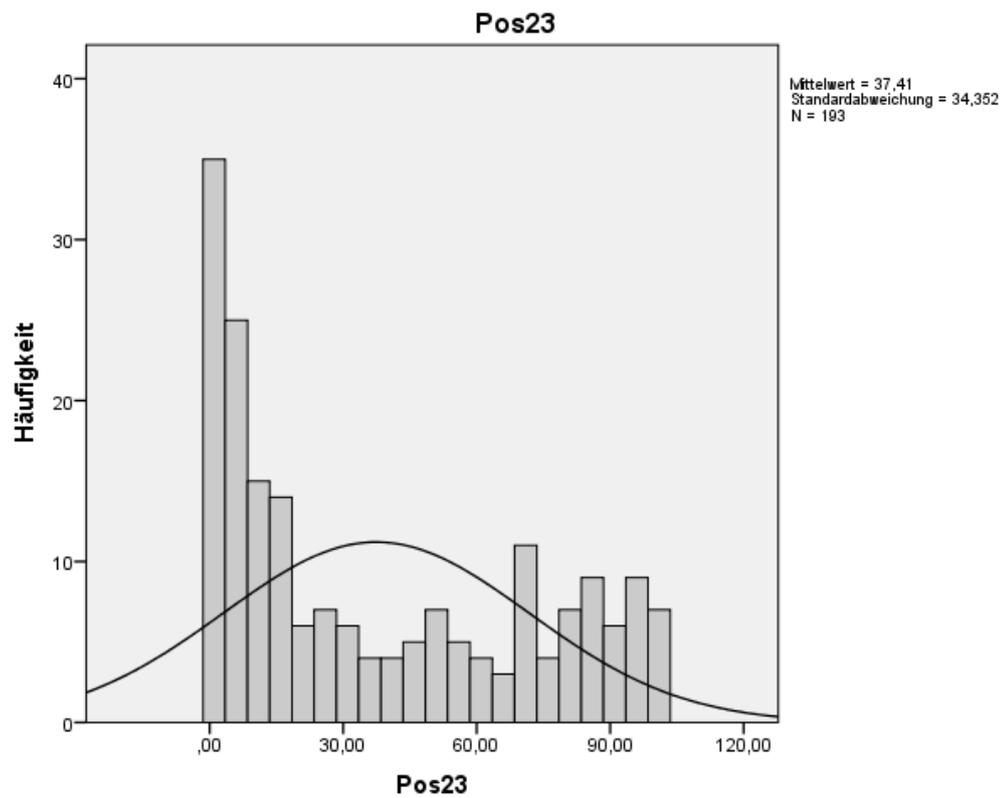


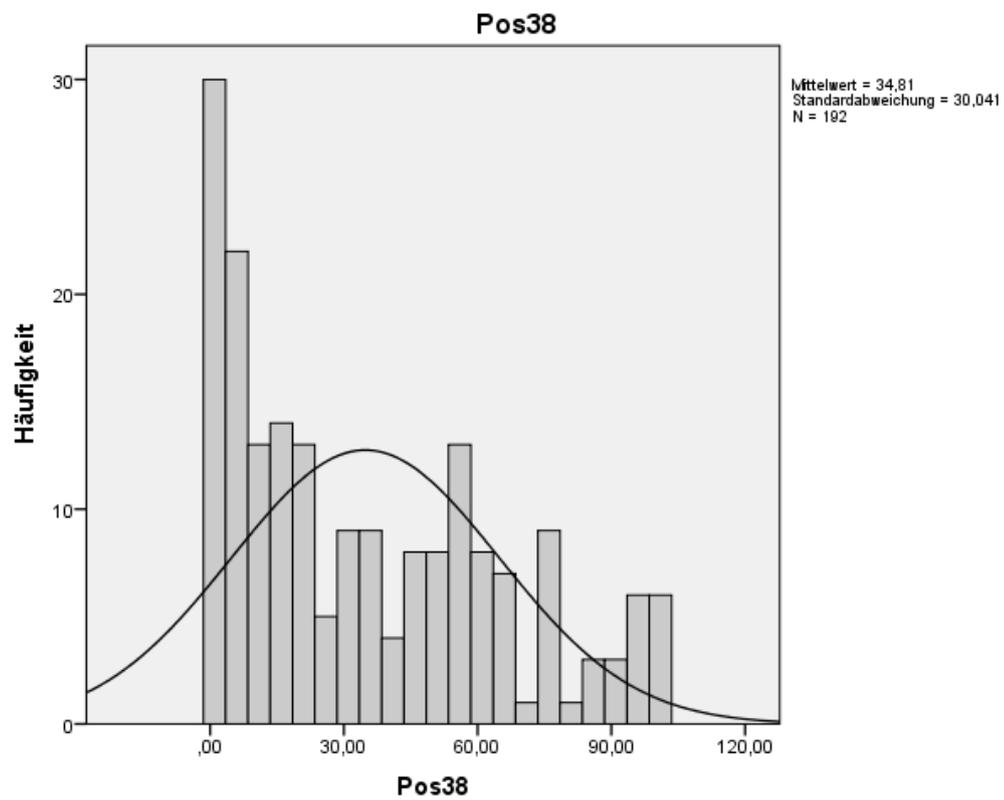
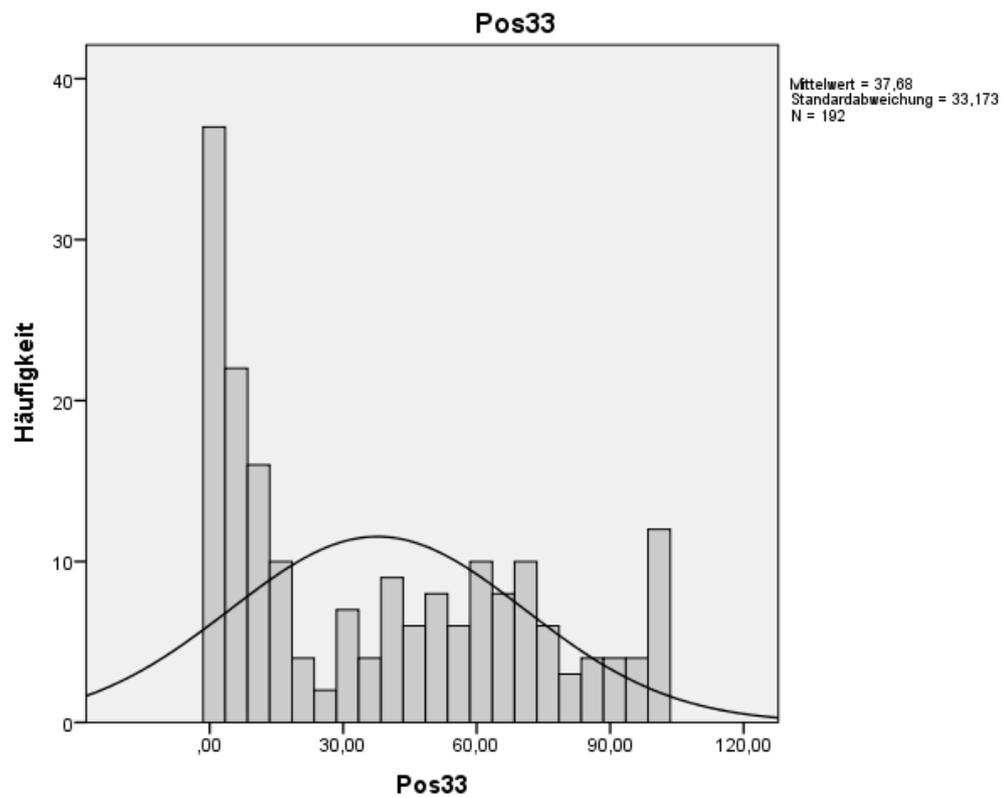


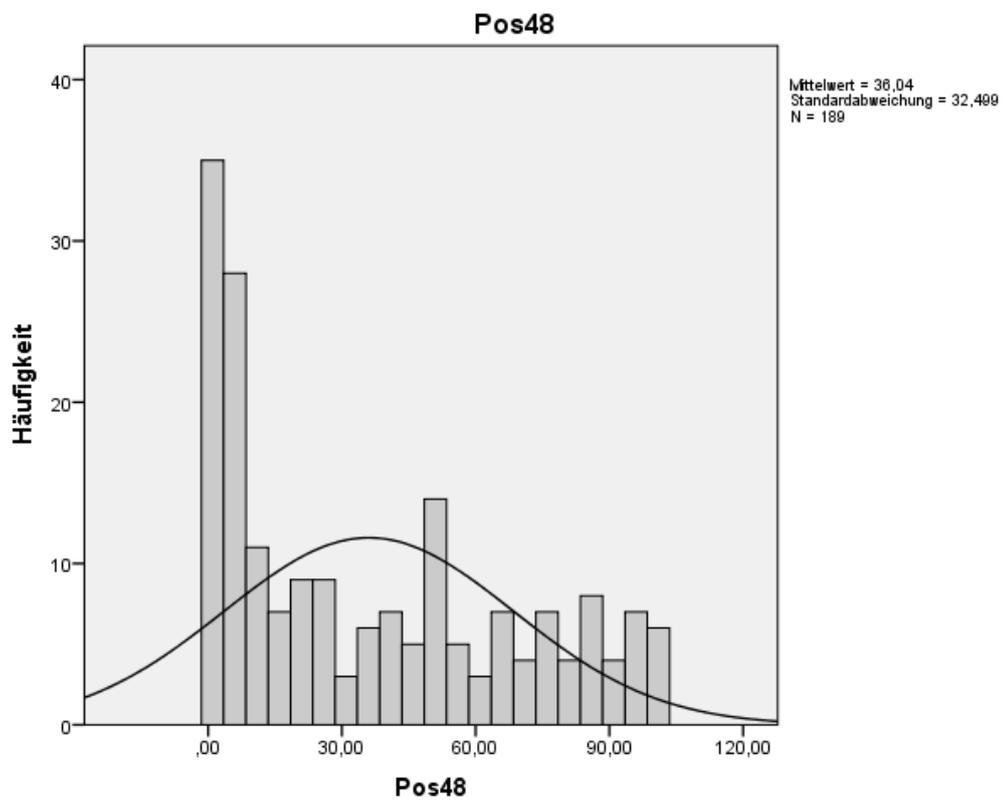
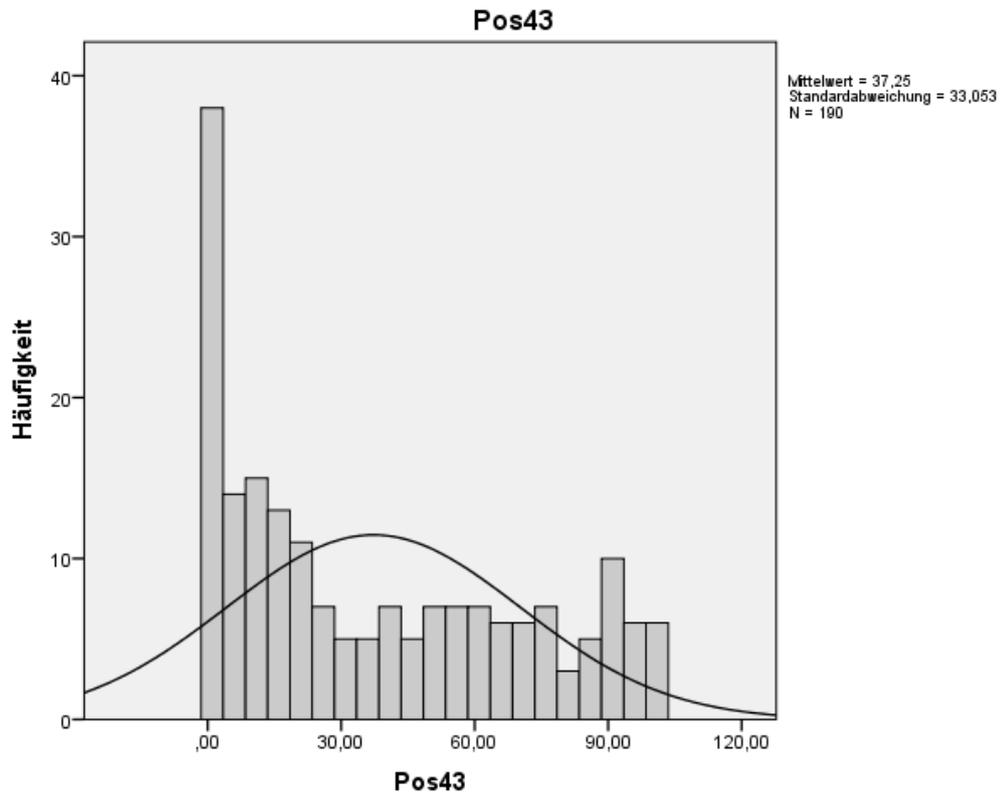


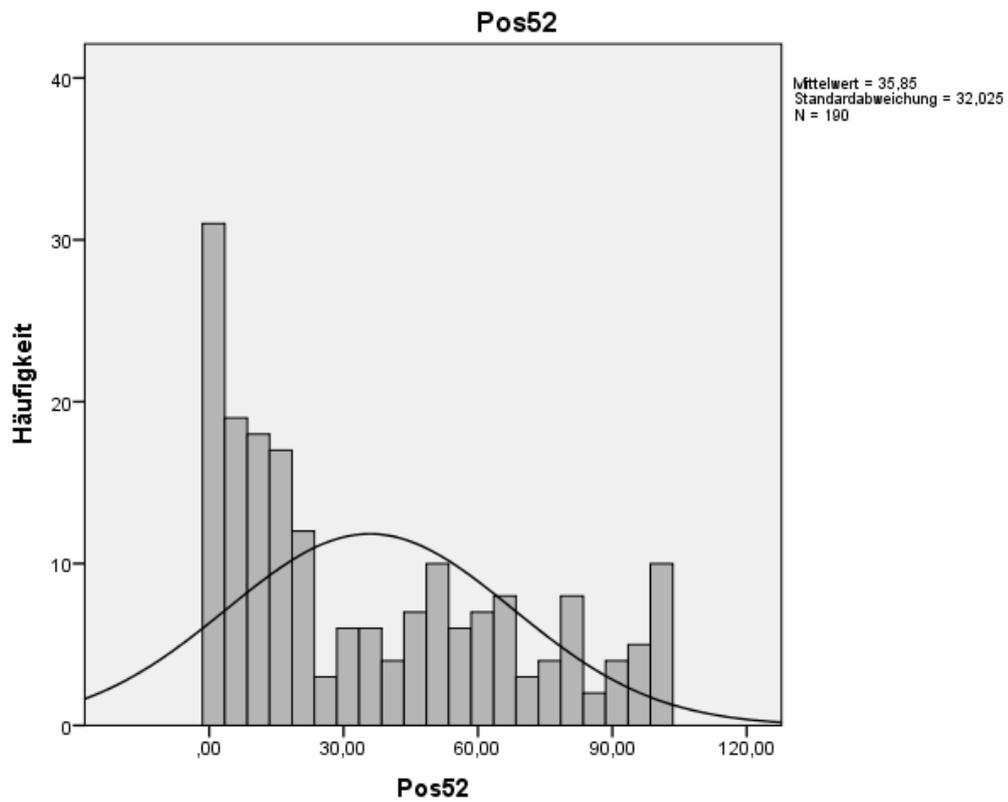
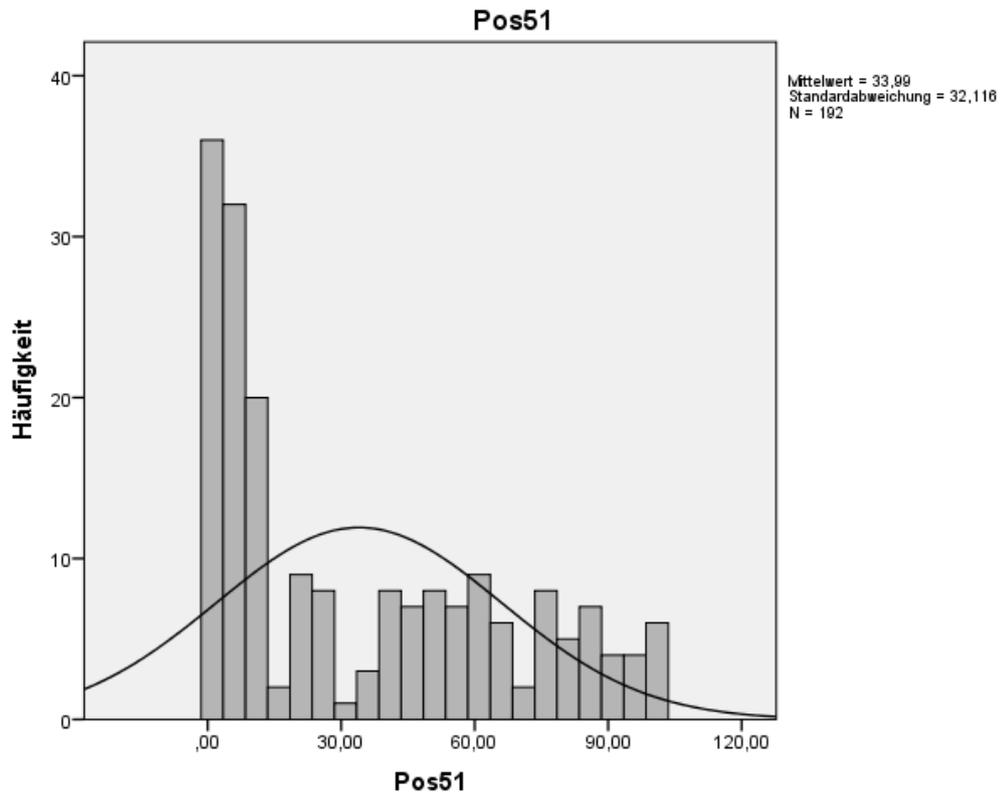


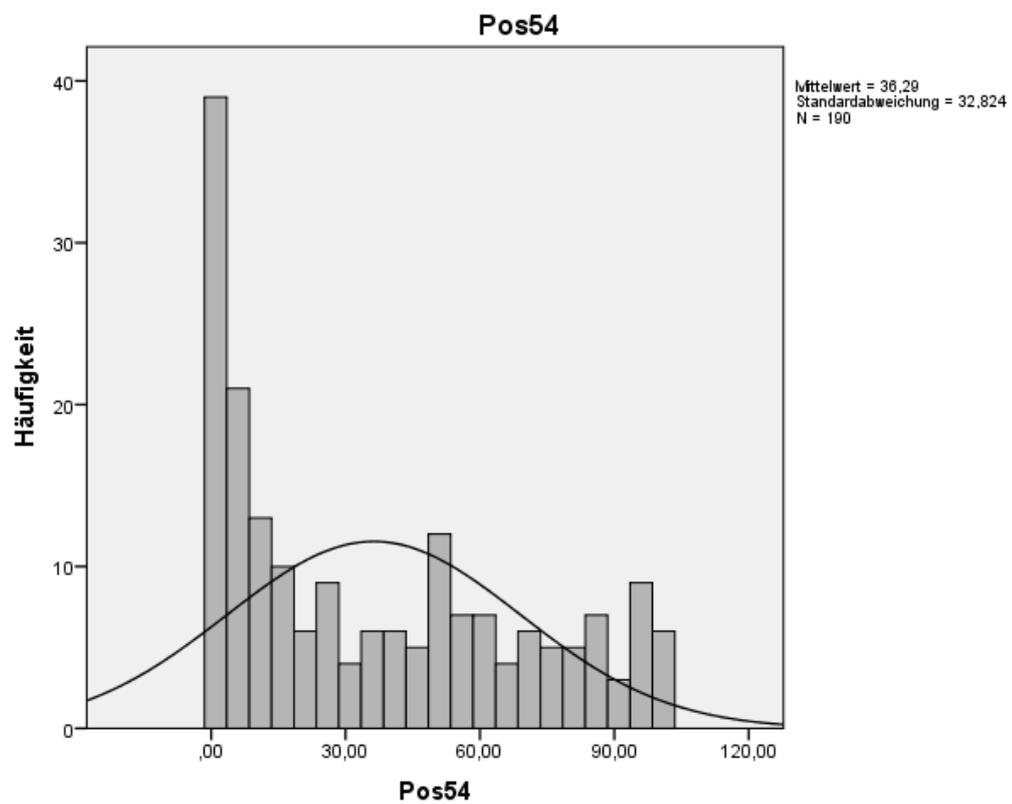
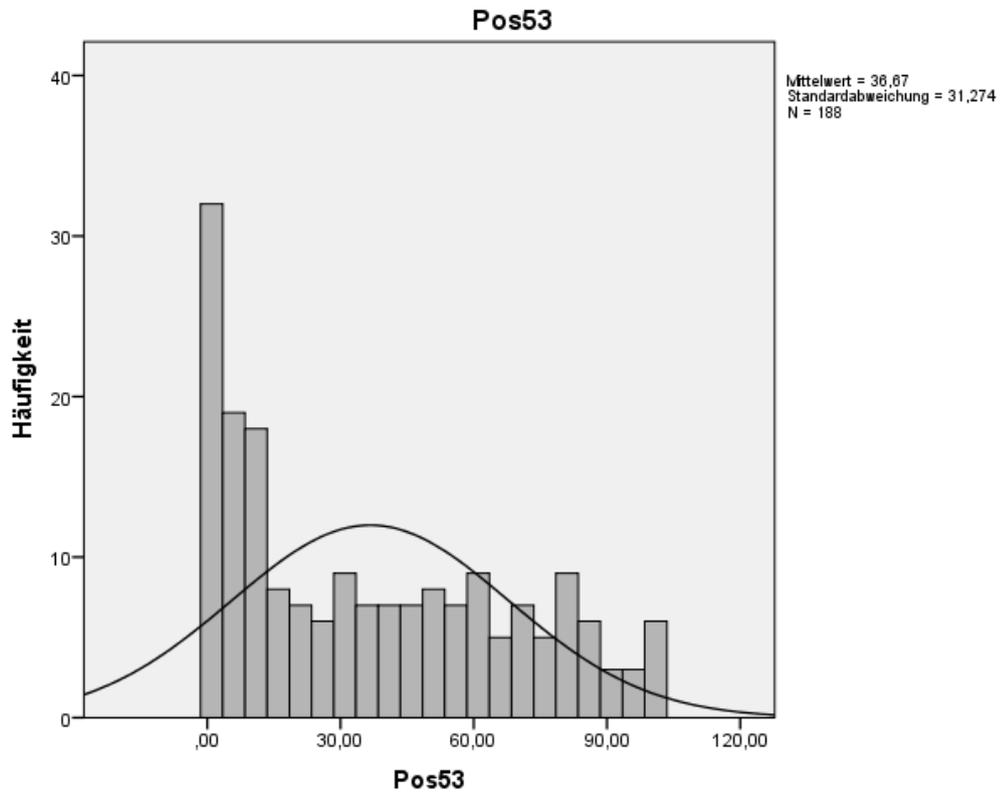


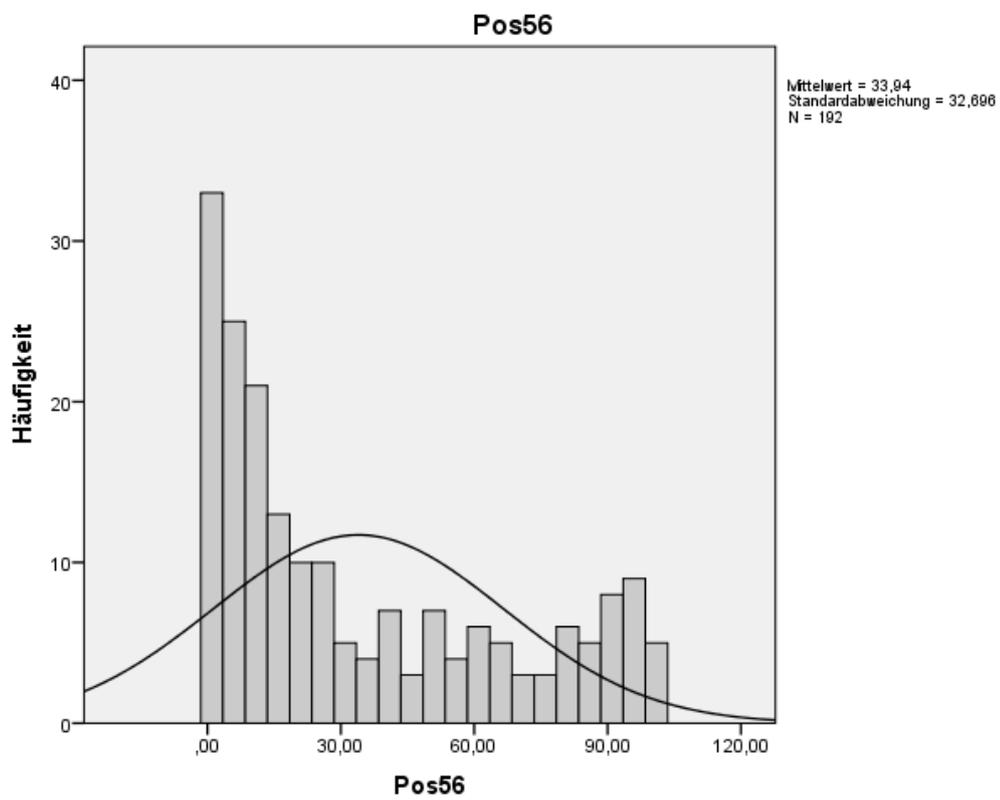
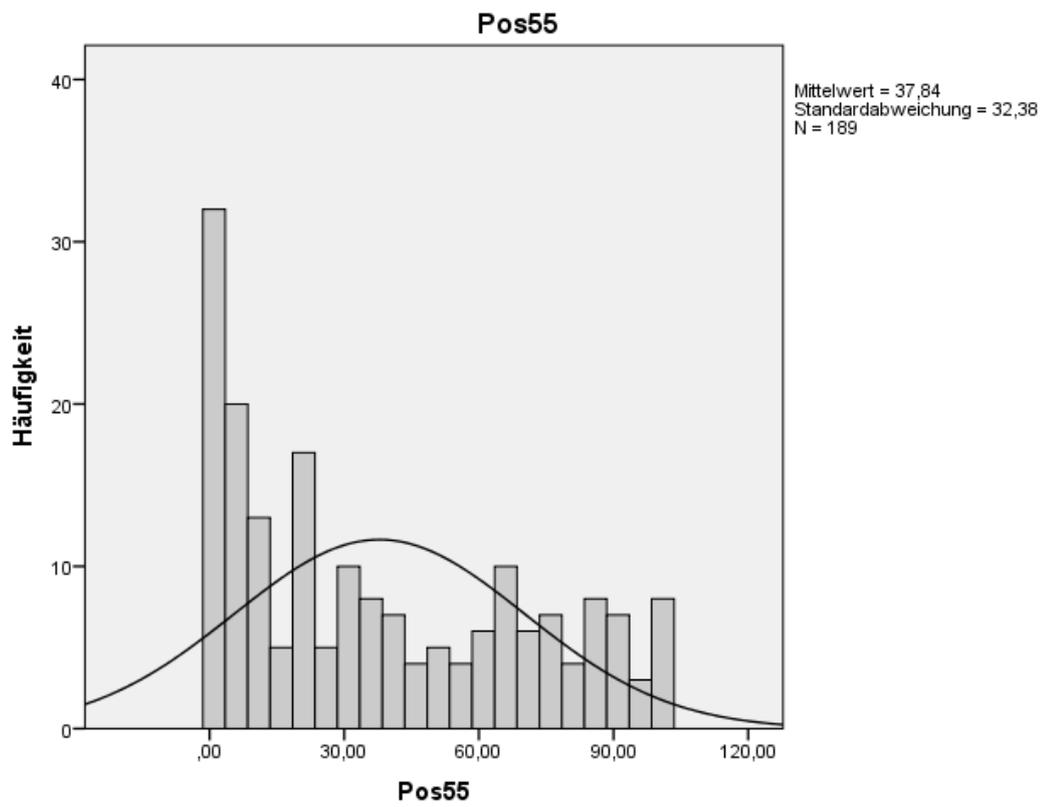


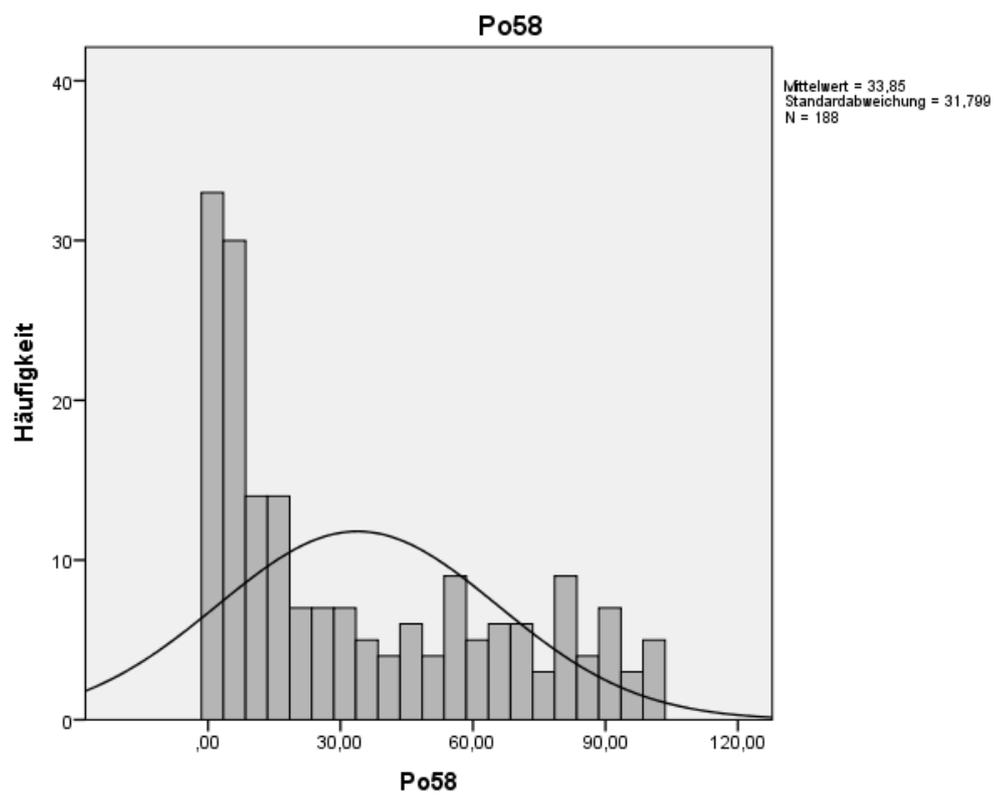
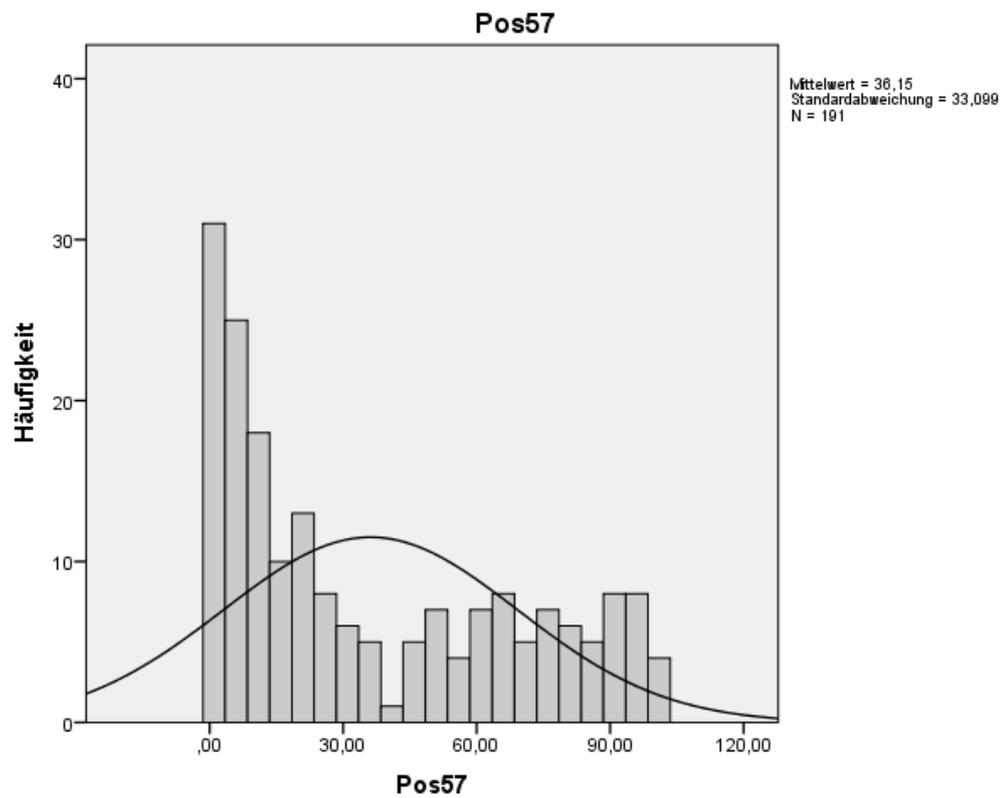


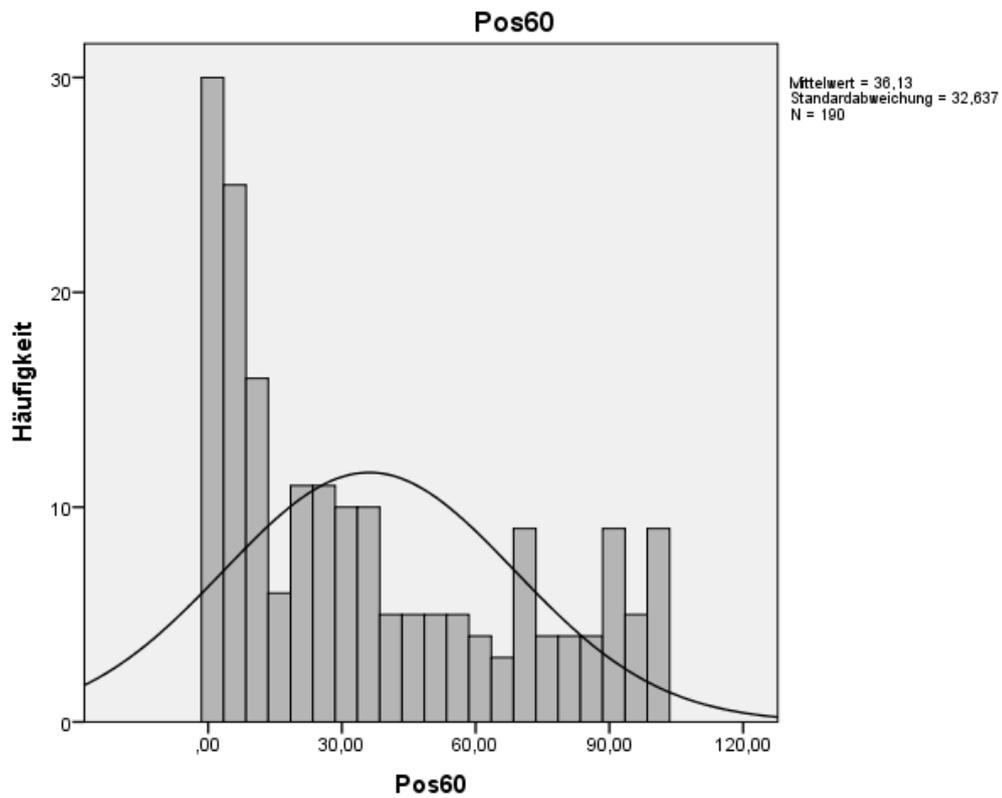
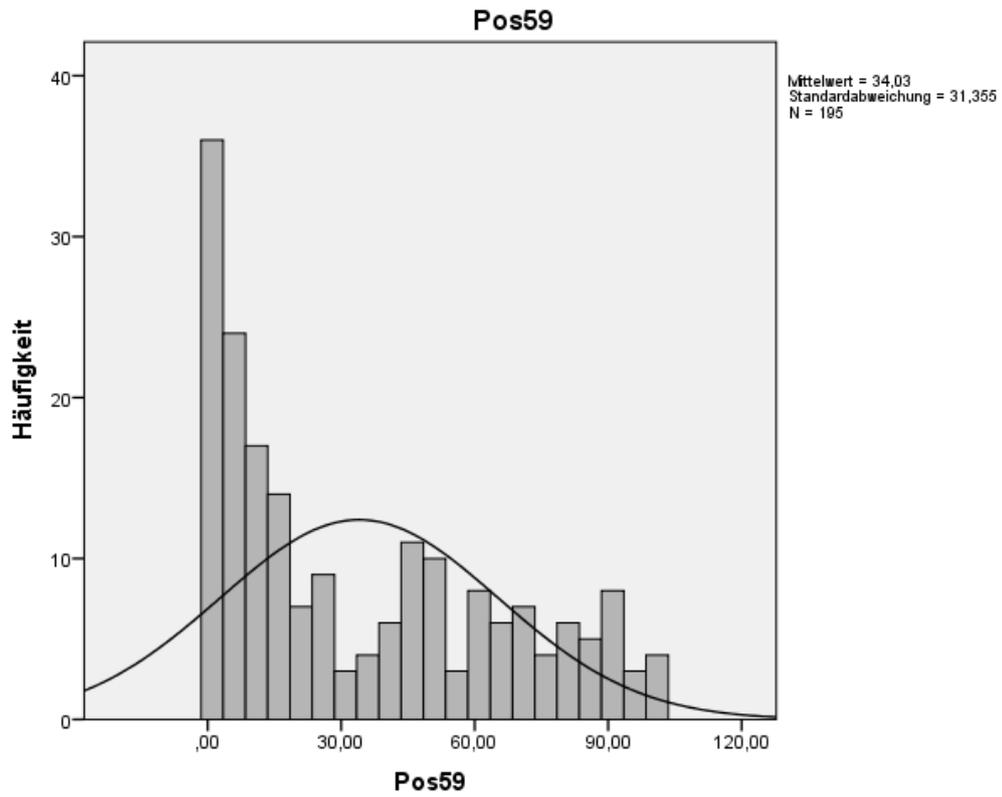


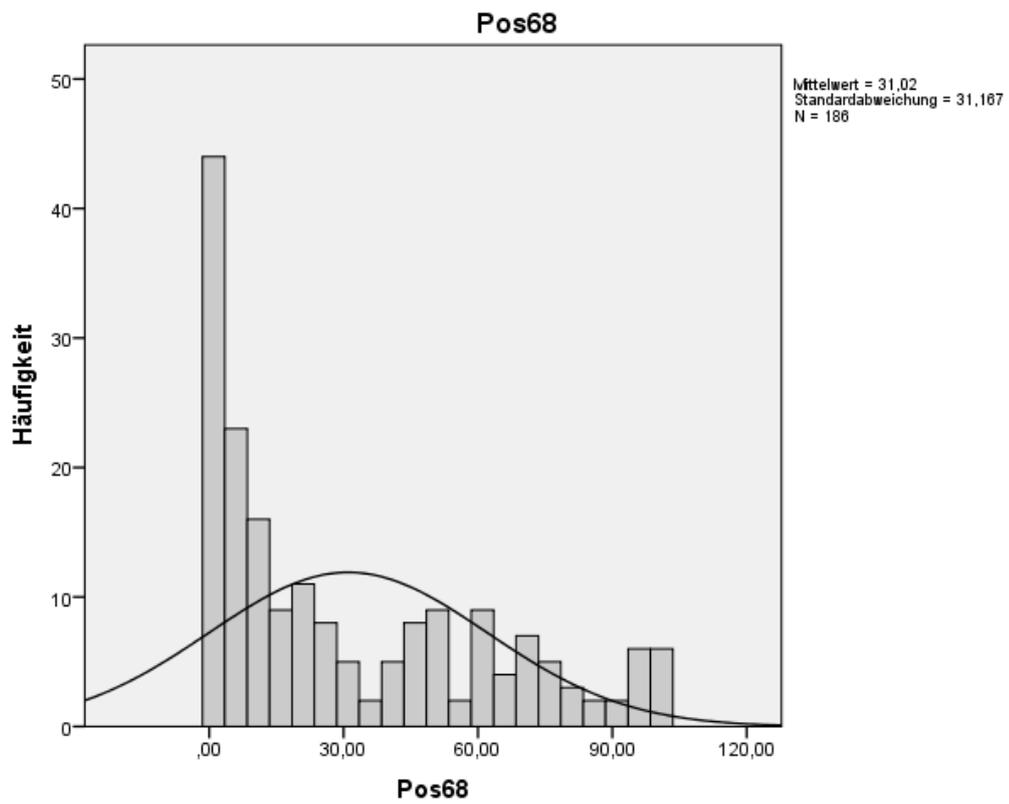
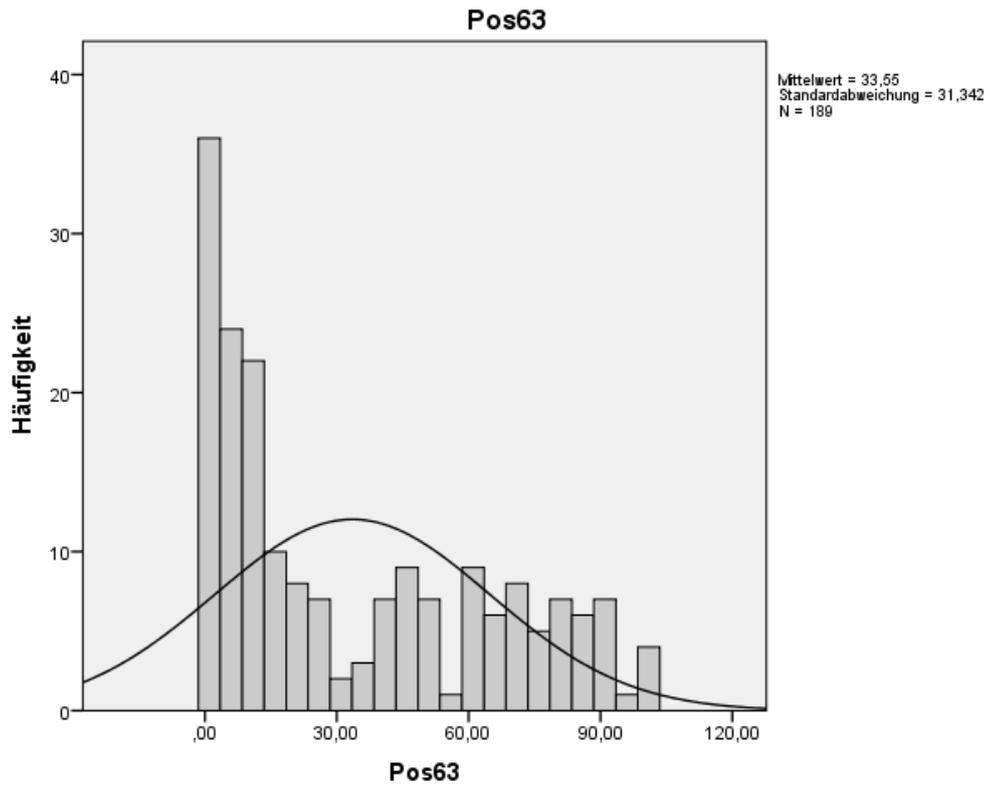


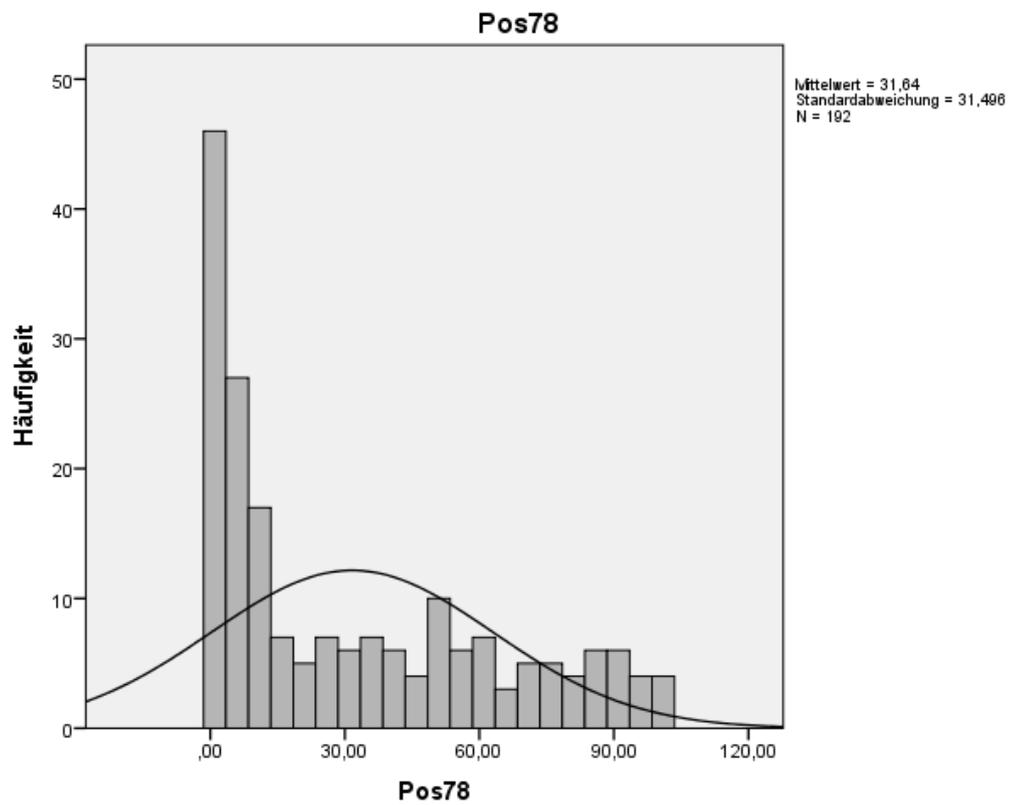
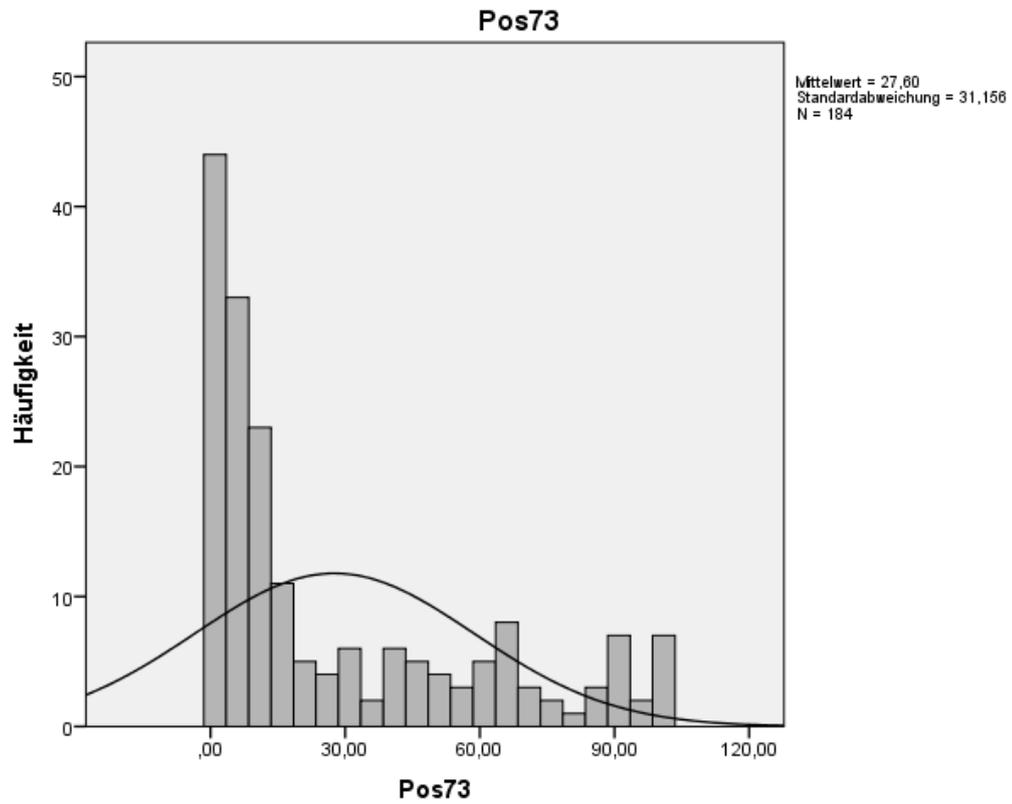


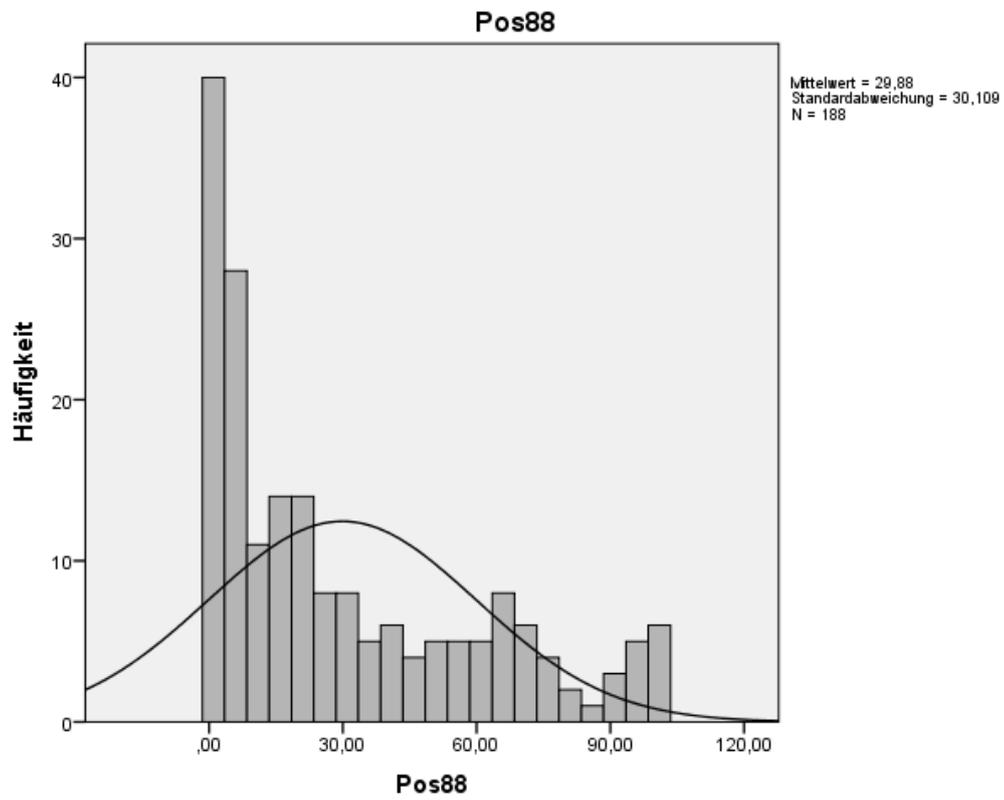
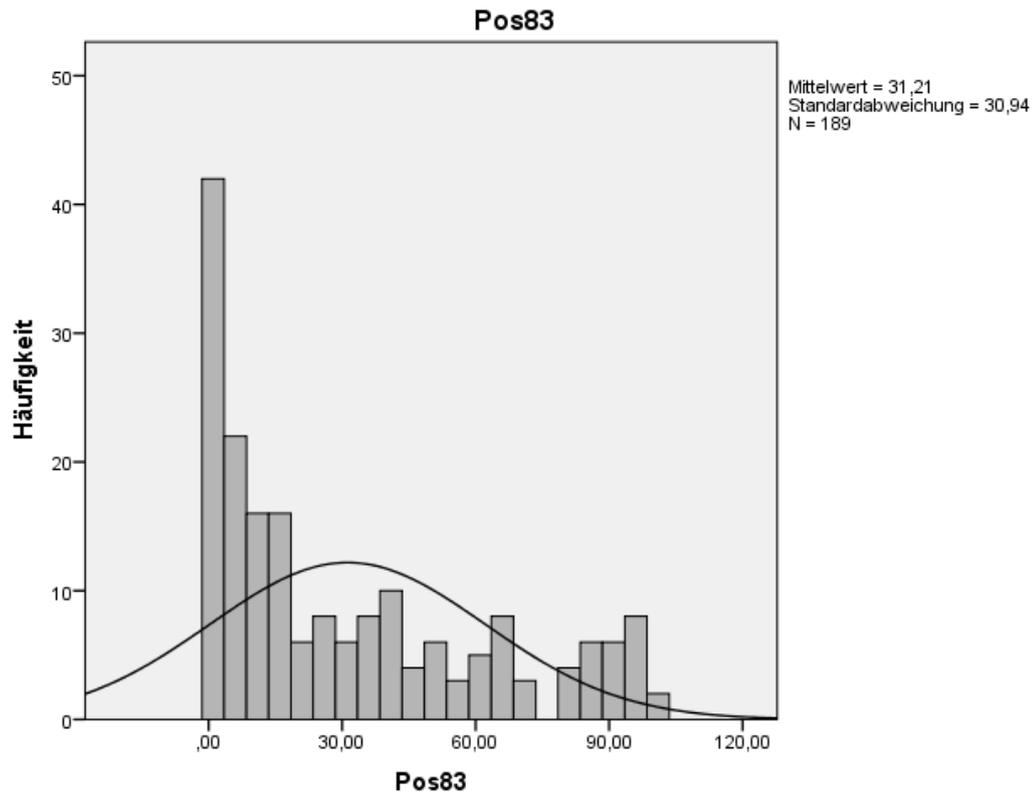


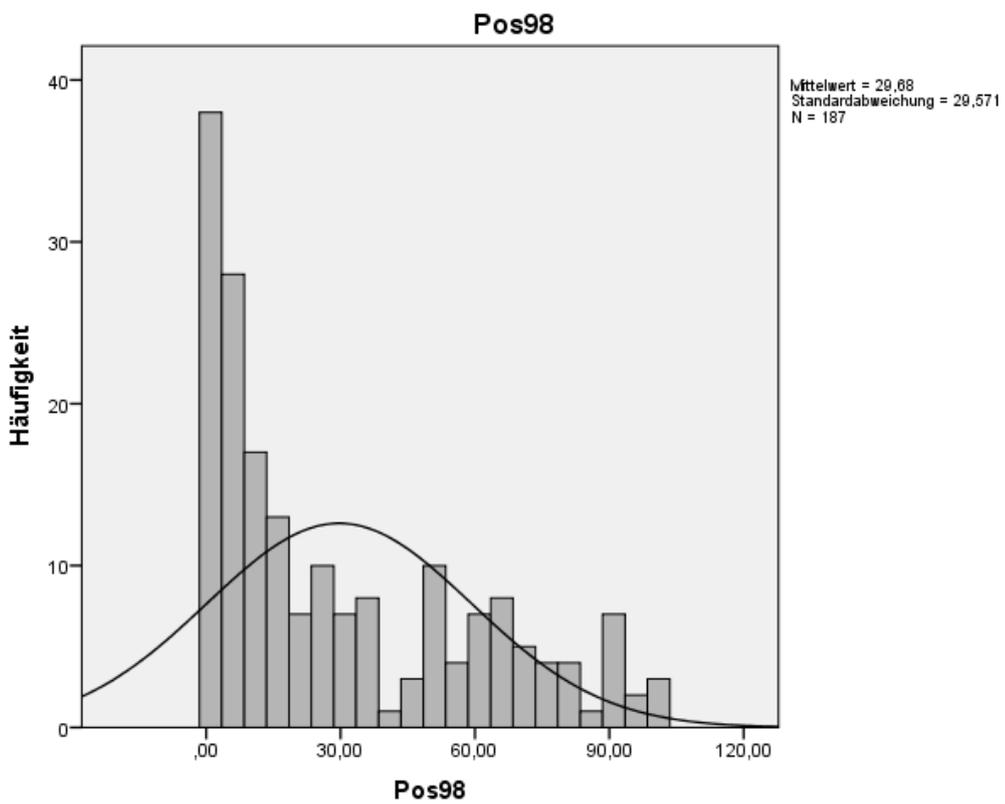
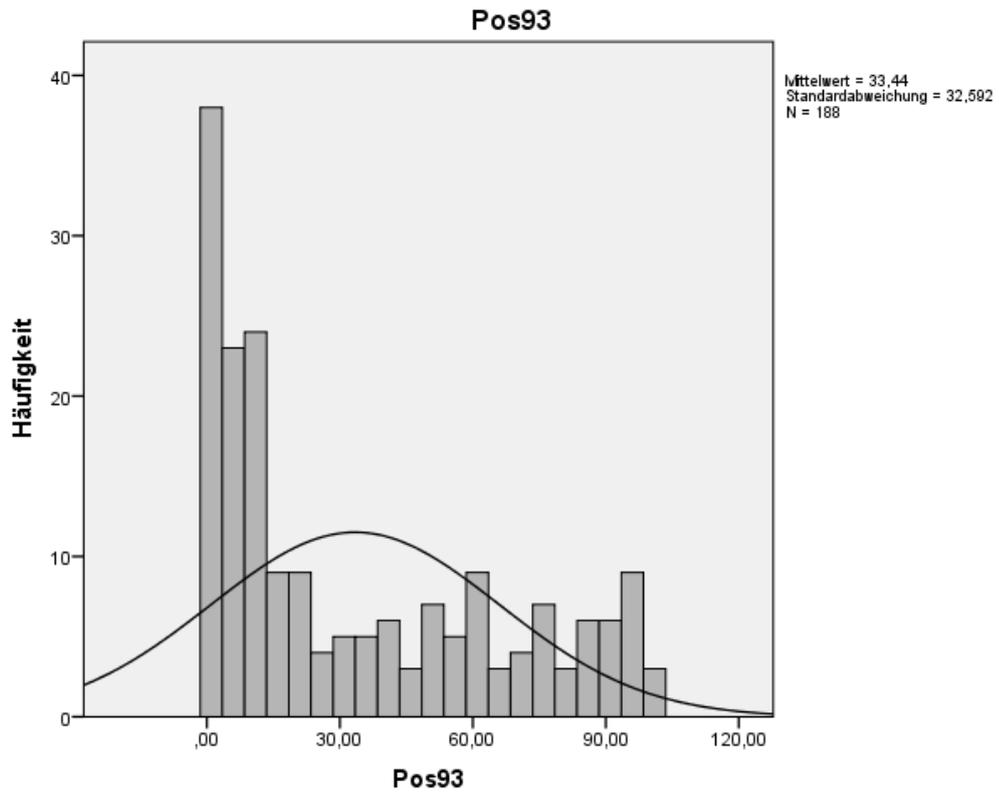












## Regression

### Eingegebene/Entfernte Variablen<sup>a</sup>

| Modell | Eingegebene Variablen | Entfernte Variablen | Methode   |
|--------|-----------------------|---------------------|-----------|
| 1      | Position <sup>b</sup> |                     | Aufnehmen |

a. Abhängige Variable: Relevanz

### Modellübersicht

| Modell | R                 | R-Quadrat | Angepasstes R-Quadrat | Standardfehler der Schätzung |
|--------|-------------------|-----------|-----------------------|------------------------------|
| 1      | ,218 <sup>a</sup> | ,048      | ,048                  | 33,133                       |

a. Prädiktoren: (Konstante), Position

### ANOVA<sup>a</sup>

| Modell       | Quadratsumme | df   | Mittel der Quadrate | F       | Sig.              |
|--------------|--------------|------|---------------------|---------|-------------------|
| 1 Regression | 376779,038   | 1    | 376779,038          | 343,205 | ,000 <sup>b</sup> |
| Residuum     | 7528878,236  | 6858 | 1097,824            |         |                   |
| Gesamtsumme  | 7905657,274  | 6859 |                     |         |                   |

a. Abhängige Variable: Relevanz

### Koeffizienten<sup>a</sup>

| Modell        | Nicht standardisierte Koeffizienten |                | Standardisierte Koeffizienten | t       | Sig.  |
|---------------|-------------------------------------|----------------|-------------------------------|---------|-------|
|               | B                                   | Standardfehler | Beta                          |         |       |
| 1 (Konstante) | 50,144                              | ,697           |                               | 71,944  | 0,000 |
| Position      | -,256                               | ,014           | -,218                         | -18,526 | ,000  |

a. Abhängige Variable: Relevanz

## Korrelation

## Korrelationen

|          |                     | Position | Relevanz |
|----------|---------------------|----------|----------|
| Position | Pearson-Korrelation | 1        | -,218 ** |
|          | Sig. (2-seitig)     |          | ,000     |
|          | N                   | 7182     | 6860     |
| Relevanz | Pearson-Korrelation | -,218 ** | 1        |
|          | Sig. (2-seitig)     | ,000     |          |
|          | N                   | 6860     | 6860     |

\*\* . Korrelation ist bei Niveau 0,01 signifikant (zweiseitig).

### Eidesstattliche Erklärung

Ich versichere, die vorliegende Arbeit selbstständig ohne fremde Hilfe verfasst und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt zu haben. Die aus anderen Werken wörtlich entnommenen Stellen oder dem Sinn nach entlehnten Passagen sind durch Quellenangaben kenntlich gemacht.

Hamburg, den .....

.....

(Unterschrift der Studierenden)