Zusammenfassung

Die vorliegende Diplomarbeit setzt sich mit der Frage der Untersuchung verschiedener Visualisierungstechniken auseinander. Um dem Leser einen kontinuierlichen Einstieg für das Thema zu ermöglichen, wurden am Anfang der Arbeit verschiedene Arten von empirischen Forschungsmethoden vorgestellt und ein Überblick über Geschichte des Eye-Trackings gegeben.

Weiterhin beschäftigt sich die Arbeit mit den Designfragen einer Vergleichsstudie im Visualisierungsbereich. Verschiedene Aspekte und mögliche Störfaktoren bei der Vorbereitung, der Durchführung und der Auswertung einer Studie werden in diesem Kontext beleuchtet.

Einen Aufgabenabschnitt der Arbeit bildet die Durchführung einer Eye-Tracking-Studie in der Graph- und Hierarchievisualisierung. Im Rahmen eines kontrollierten Experiments wurden drei Darstellungsarten von Baumdiagrammen verglichen. Die Aufgabe der Probanden bestand darin, den kleinsten gemeinsamen Vorfahrknoten aller rot markierten Blätter zu finden.

Im letzten Abschnitt der Arbeit werden die Vergleichstudie und ihre Ergebnisse präsentiert.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

1.1 Problemstellung

Die technologische Entwicklung durchdringt mit steigendem Tempo alle Bereiche des menschlichen Umfelds. Das hat zur Folge, dass Daten und Zahlen eine immer wichtigere Rolle im Alltag spielen. Dabei sind Visualisierungstechniken zu einem vielseitigen und mächtigen Instrument geworden, die mit Hilfe von Personalcomputern den Benutzer beim Umgang mit den großen Datensätzen unterstützt.

Das Hauptproblem bei dieser Entwicklung besteht darin, dass die Menge der zu verarbeitenden Daten ununterbrochen wächst: Einerseits werden naturwissenschaftliche Messtechnologien präziser, andererseits erhöht sich die Zahl der wissenschaftlichen Publikationen sowie erhobener Wirtschaftsdaten. Diese Informationsmenge wird nach wie vor auf dem Computerbildschirm dargestellt, dessen Größe jedoch weitgehend unverändert bleibt (vgl. [1]).

Sowohl dieser Aspekt als auch die Grenze der Lesbarkeit bestimmen die darstellbare Datenmenge, so dass moderne Datenbanken in ihrem vollen Umfang kaum noch in numerischer Form darstellbar sind. Für deren repräsentative Darstellung ist es nicht einfach, einen relevanten Ausschnitt der Daten auszuwählen. Dabei besteht das Risiko, zu wenige oder falsche Datensätze zu wählen. Außerdem nehmen die Menschen bei größeren numerischen Datendarstellungen aufgrund der begrenzten kognitiven Möglichkeiten nur einen kleinen Ausschnitt wahr.

Große Datenmengen können jedoch durch visuelle Transformation in ihrer Gänze präsentiert werden, so dass neue und interessante Einsichten entstehen. Durch Visualisierung von Daten werden oft unerwartete Strukturen, Beziehungen und Trends erkennbar, die sonst in einer Menge von Zahlen unentdeckt bleiben (vgl. [1]).

Allerdings sollten die Vorzüge von Visualisierungen nicht von vornherein ohne aussagekräftige Ergebnisse von empirischen Untersuchungen angenommen werden. Claus Arnold [1] hat in seiner Studie zwei Suchmaschinen mit Visualisierungsaufsätzen (Kartoo und Webbrain) mit der traditionellen Trefferdarstellung als eine Liste (Google) verglichen und festgestellt, dass die WWW-Suchmaschinen mit dem Aufsatz von Visualisierungskomponenten schlechtere Benutzbarkeit aufweisen. Dieses Ergebnis erklärt, warum sich solche Anwendungen bisher nicht verbreiten konnten und zeigt, dass das Prinzip „Ein Bild sagt mehr als tausend Worte“ in jedem konkreten Fall überprüft werden sollte. Diese Anforderung ist jedoch nicht leicht umzusetzen. Zwar gibt es eine Reihe von Studien, die verschiedene Visualisierungstechniken vergleichen und untersuchen [5, 6, 11, 19], es gibt aber keine allgemeine systematische und ausführliche Arbeit, die das Design einer Vergleichsstudie für den Visualisierungsbereich vorschreibt und alle möglichen Facetten bei solch einer Problemstellung beleuchtet.

Die vorliegende Arbeit bestrebt diese Lücke zu schließen und zusätzlich einen Beitrag zur empirischen Forschung im Bereich von Hierarchievisualisierung zu leisten. Zunächst wird ein Überblick über empirische Forschungsmethoden insbesondere über Eigenschaften der kontrollierten Experimente gegeben. Danach folgt die Beschreibung der Vor- und Nachteile des Eye-Tracking-

Verfahrens. Da Eye-Tracker-Geräte heutzutage als leicht bedienbar, mächtig und präzise gelten, sind sie praktisch unverzichtbar bei der Durchführung von empirischen Untersuchungen im Visualisierungsbereich. Ein Eye-Tracker-Gerät wurde in der im Rahmen dieser Diplomarbeit ausgeführten Studie verwendet, um die Blickbewegungen der Probanden zu verfolgen und aufzuzeichnen.

Im nächsten Abschnitt der Arbeit werden die allgemeinen Design-Vorschläge für eine Vergleichsstudie im Visualisierungsbereich vorgestellt. Die Durchführung einer Benutzerstudie ist mit vielen Problemen und Hindernissen verbunden. Diese sollten beim Entwurf und der Durchführung der Studie sowie bei der Auswertung der Ergebnisse beachtet werden, damit die Studie möglichst reibungslos verläuft und ihre Resultate nicht verfälscht werden. Danach werden die einzelnen Schritte und verschiedene Aspekte der durchgeführten Vergleichsstudie beschrieben. Am Ende der vorgelegten Arbeit werden die Ergebnisse der Studie präsentiert und diskutiert. Sie untersucht den Einfluss von unterschiedlichen Anordnungen der Knoten-Kanten-Diagramme auf die Richtigkeit und Geschwindigkeit bei der Lösbarkeit von Visualisierungsaufgaben, insbesondere auf die Fragestellung, welcher Knoten der kleinste gemeinsame Vorfahr einer gegebenen Menge von Blattknoten ist.

1.2 Visualisierung

Die visuelle Darstellung von Daten wird schon seit einigen Jahrhunderten verwendet. William Playfair (1759 - 1823) wird als Begründer der Visualisierung der statistischen Daten angesehen. Die erste Visualisierung dieser Art entstand im Jahr 1786 und zeigt das Verhältnis von Preisen, Gehältern und Regierungszeiten britischer Königinnen und Könige (Abb. 1) (vgl. [1]).

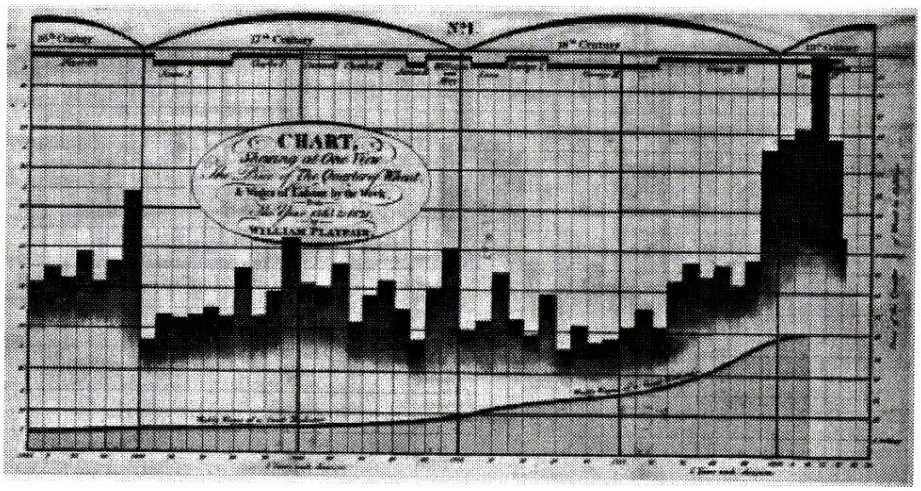

Abbildung 1: Eine der ersten Visualisierungen von William Playfair [1]

Bertin entwickelte erste theoretische Konzepte über Graphik, in denen die Grundelemente von Diagrammen definiert und Gestaltungsrichtlinien festgelegt wurden.

Sehr ertragreich waren die Forschungen im Bereich graphischer Darstellungen von Statistiken. Tukey verwendete Bilder [29] und Cleveland/McGill dynamische Graphiken [7], damit die

Präsentation von Daten übersichtlicher wird. 1983 entstand die Theorie von Tufte, die die Darstellungsgrenzen großer Datenmengen in statistischen Graphiken untersucht (vgl. [1]).

Die rapide Steigerung von Hardwareleistung ermöglichte den Naturwissenschaftlern bei ihrer Forschung immer aufwändigere Berechnungen durchzuführen und die Ergebnisse graphisch zu präsentieren. In den 80er Jahren wurden hauptsächlich konkrete physische Daten aus dem meteorologischen und astronomischen Bereich visualisiert.

Die Darstellung großer Mengen von Satellitendaten war das Hauptthema auf der ersten IEEE Visualization Conference für Geologen, Physiker und Computerwissenschaftler, die 1990 in San Francisco stattfand. Danach wurden Visualisierungsverfahren ebenfalls für Optimierung von Benutzerschnittstellen eingesetzt, um Dokumente und Datenbanken mit multivariablen Daten übersichtlicher darzustellen (vgl. [1]).

1.3 Visual Analytics

Das Problem des Umgangs mit den ständig wachsenden Datenmengen wird zusätzlich durch das Information-Overload-Phänomen verschärft: Die Steigerung der Hardwareleistung drückt sich hauptsächlich durch die schnelle Erweiterung des Speicherplatzes aus. Dabei zeigen die Rechenleistung und Auswertungsmöglichkeiten aber viel langsamere Entwicklungen.

Laut einer Studie des Marktforschungsunternehmens International Data Corporation wird im Jahr 2011 die digital produzierte Datenmenge etwa 1.800 Exabyte betragen und sich im Vergleich zum Jahr 2006 verzehnfacht haben. Doch im Schnitt 80% der gespeicherten Daten werden nie verwendet. Die Suche nach den wirklich relevanten Informationen wird in dem sich stets ausbreitenden Datenozean immer komplexer und kostenintensiver (vgl. [30]).

Der Nachteil der manuellen Datenauswertungsverfahren zur Informationsgewinnung besteht in der hohen Fehlerwahrscheinlichkeit. Auch Data-Mining-Verfahren, bei denen durch die Interaktion von Mensch und Maschine Modelle von Regularitäten und Zusammenhängen in den Daten gebildet werden, stoßen bei großen Datenmengen schnell an ihre Grenzen. Die wichtigsten Gründe dafür sind die Komplexität von Algorithmen und die Datenqualität. In vielen Fällen müssen bei Verwendung dieser Verfahren die Algorithmen für einen Datenbestand speziell parametrisiert werden. Dafür wird aber das Verständnis für die Funktionsweise der Algorithmen vorausgesetzt, was in der Praxis nur dem engen Kreis von Experten zugänglich ist.

Im Arbeitsalltag von vielen auch nichttechnischen Berufen müssen die Datenmengen einem systematischen Erkenntnisprozess unterzogen werden. Dabei soll ausgehend von Daten und Informationen, Wissen aus den Daten gewonnen werden. Für solch einen Auswertungsprozess müssen neue Verfahren zur explorativen Datenanalyse entwickelt werden. Als eine Lösung dafür wird die visuelle, explorative Datenanalyse (Visual Analytics) gesehen (vgl. [30]).

Visual Analytics ist eine interdisziplinärere Methode. Sie verbindet die Stärken der automatischen Datenanalyse mit der Fähigkeit des Menschen, schnell Muster oder Trends visuell zu erfassen. Dieser Ansatz ist sehr effizient bei der Gewinnung von Erkenntnissen aus extrem

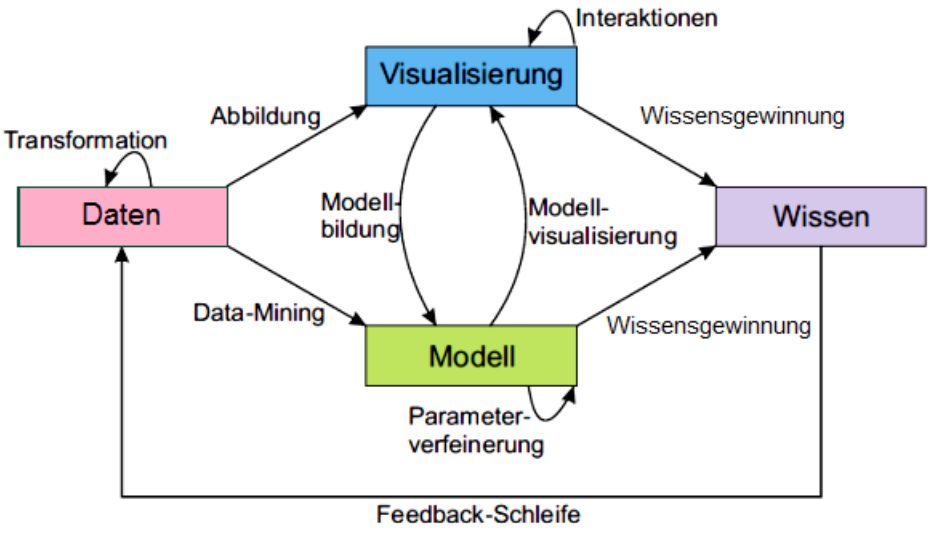

Abbildung 2: Der Visual-Analytics-Prozess [12]

großen und komplexen Datensätzen. Mit Hilfe von geeigneten Interaktionsmechanismen können Daten visuell exploriert und neues Wissen gewonnen werden [13]. Jim Thomas beschreibt ausführlich diese Methode in seinem Buch "Illuminating the Path" [27].

Das Ziel von Visual Analytics besteht darin, den Mensch mit seiner Flexibilität, Kreativität und seinem Allgemeinwissen in den Analyseprozess einzubinden. Wie bei Data Mining entsteht so die Ergänzung zu algorithmischen Verfahren. Auf diese Weise ist es möglich, das Information-Overload-Problem in einen Vorteil umzukehren.

In Abbildung 2 ist die Übersicht über die Komponenten des Visual-Analytics-Prozess dargestellt. Der Prozess besteht aus verschiedenen Zuständen (dargestellte Blöcke) und Transformationen (Pfeile) (vgl. [30]).

In vielen Visual-Analytics-Szenarien muss zuerst die Integration der heterogenen Datenquellen durchgeführt werden. Ergibt diese Transformation aussagekräftige Dateneinheiten, so besteht im nächsten Schritt die Möglichkeit, zwischen einer visuellen oder einer automatischen Analysemethode auszuwählen. Schon nach dieser Phase kann das gewünschte Wissen herausgefiltert werden. Ist dies nicht der Fall, sind zusätzliche Benutzerinteraktionen wie zum Beispiel das Zoomen in verschiedenen Bereichen der Darstellung notwendig, um weitere Informationen zu bekommen. Auf diese Weise könnten Details zu einer Datenmenge betrachtet werden.

Die Visualisierungsmodelle, die durch Interaktion entstanden und verändert wurden, können zur automatischen Analyse wiederverwendet werden. In dieser Tatsache besteht der Unterschied des Visual Analytics zu klassischen Informationsvisualisierungen. Dieselben Visualisierungsmodelle könnten auch mit Hilfe von Data-Mining-Techniken aus den originalen Daten erstellt werden (vgl. [30]).

Ein erstelltes Modell lässt sich außerdem durch Interaktion mit den automatischen Analysemethoden verbessern, indem Parameter verändert oder andere Arten von Analysealgorithmen

ausgewählt werden. Um das Modell zu verifizieren, kann wieder eine Modellvisualisierung verwendet werden.

Der wichtigste Vorteil der Visual-Analytics-Methode besteht in dieser Möglichkeit des Wechsels zwischen visuellen und automatischen Werkzeugen. Dadurch ist eine kontinuierliche Verfeinerung und Verifikation von vorläufigen Ergebnissen möglich. Da irreführende Resultate in einem frühen Zwischenschritt entdeckt werden können, sind auch die Endergebnisse vertrauenswürdiger und zuverlässiger als bei den anderen Methoden der Datenexploration. Die Feedback-Schleife im Prozess dient der Sicherung des durch die Analyse gefundenen Wissens (vgl. [30]).

Es ist offensichtlich, dass alle vier in Abbildung 2 dargestellten Zustände eine wichtige Rolle im gesamten Visual-Analytics-Prozess spielen. Die Zustände sowie die Übergänge zwischen ihnen müssen gründlich entworfen und gut realisiert werden.

Da das menschliche visuelle Wahrnehmen jedoch individuell sehr stark variiert, könnten besonders viele Fragen bei der Entwicklung der Visualisierungskomponenten entstehen. Es gibt beispielsweise keine Garantie, dass Visualisierungen von Benutzern so interpretiert und verstanden werden, wie dies deren Autoren erwarten. Der weitere Problemaspekt ist die Frage, ob eine konkrete Visualisierung für den gegebenen Datensatz die beste ist. Stellt sie die Abhängigkeiten gut dar? Setzt sie die richtigen Schwerpunkte? Hebt sie das Wichtigste hervor?

Diese und ähnliche Fragen lassen sicht nicht ohne gezielte Untersuchungen beantworten. Damit aber beim Vergleich von Visualisierungen vertrauenswürdige Ergebnisse erzielt werden können, sind ein großer Aufwand und systematische Vorgehensweise notwendig. Ein Überblick über die empirischen Forschungsmethoden, die für die Untersuchung der Qualität von Informationsvisualisierungen hilfreich sein können, wird im Kapitel 2 gegeben.

2 Empirische Forschungsmethoden

Es existieren verschiedene Arten von wissenschaftlichen Untersuchungsmethoden. Meistens liefern aber nur Kombinationen von mehreren Methodentypen ausreichende Untersuchungsergebnisse (vgl. [17, 23]). Eine große Gruppe von Ansätzen bilden die empirischen Forschungsmethoden. Sie spielen in der Forschung eine große Rolle und sind effektiv bei der Untersuchung einer bestimmten Frage oder bei der Suche nach Ursache und Wirkung (vgl. [16]).

Unter „**Empirie**“ werden Erkenntnisse verstanden, die auf Erfahrungen beruhen. „Empirische Beobachtungen oder Aussagen beziehen sich auf Wahrnehmungen und/oder sind von solchen abgeleitet“ [22]. Aber allein das Vorhaben, „etwas empirisch untersuchen zu wollen“, reicht für die verlässliche wissenschaftliche Arbeit nicht aus. Für die Durchführung empirischer Forschung sind umfassende Kenntnisse empirischer Forschungsmethoden notwendig (vgl. [3]). Im Folgenden wird ein Überblick über diese Methoden sowie über deren relevante Eigenschaften gegeben.

2.1 Relevante Eigenschaften von empirischen Forschungsmethoden

Der Wert einer empirischen Forschungsmethode wird daran gemessen, inwieweit sie den inhaltlichen Erfordernissen einer Untersuchung gerecht wird. Eine Forschungsmethode ist niemals für sich genommen gut oder schlecht (vgl. [3]). Mit Hilfe von **Qualitäts-** und **Eignungsaspekten** lässt sich feststellen, ob eine Methode im konkreten Fall vorteilhafter als eine andere ist. Die wichtigsten Aspekte sind

- **Zeit:** In manchen Forschungsfällen werden die Ergebnisse möglichst schnell benötigt, etwa wenn sich eine ansteckende Krankheit ausbreitet, von der noch nicht so viel bekannt ist. Dann müssen viele Menschen möglichst schnell befragt werden, um schleunigst brauchbare Einsichten zu bekommen.

- **Aufwand und Kosten:** Ist der zeitliche Aspekt weniger relevant, so wird die finanzielle Frage oftmals entscheidend für die Auswahl einer Forschungsmethode sein (vgl. [20]).

- **Verlässlichkeit der Resultate:** Sie bedeutet die minimale Wahrscheinlichkeit, dass die Resultate der Forschung durch eine fehlerhafte Planung oder Durchführung verfälscht wurden und schließt die Irrtümer bei der Interpretation der Ergebnisse praktisch aus. Vor allem bei Forschungsmethoden mit weniger Kontrolle über den Beobachtungsvorgang gehört die mangelnde Verlässlichkeit zu den Nachteilen (vgl. [20]).

- **Beschreibbarkeit und Verstehbarkeit:** Bei guter Beschreibbarkeit lässt sich die Forschung leicht und präzise dokumentieren, so dass andere Wissenschaftler sie vollständig verstehen können. Die mangelnde Beschreibbarkeit ist meistens auf Komplexität der Forschung oder Komplexität der Umgebung, in der sie stattfand zurückzuführen (vgl. [20]).

- **Reproduzierbarkeit der Forschung und der Resultate:** Die Reproduzierbarkeit kennzeichnet den Aufwand und die Schwierigkeit, die bewältigt werden müssen, um eine fremde Forschung zu wiederholen. Sie geht über Verlässlichkeit und Beschreibbarkeit hinaus, da es praktisch unmöglich ist, eine bestimmte Situation zu reproduzieren, selbst wenn diese genau beschrieben und fehlerfrei untersucht wurde (vgl. [20]).
- **Verallgemeinerbarkeit der Resultate**: Dieser Aspekt ist maßgeblich für die Relevanz der Untersuchung und hängt von der Anzahl der Fälle ab, auf die sich das Ergebnis direkt oder mit bestimmten Modifikationen übertragen lässt (vgl. [20]).

Die empirische Forschung unterscheidet sich vom alltäglichen Erkenntnisgewinn dadurch, dass Erfahrungen gesammelt und dokumentiert werden. Das Problem besteht darin, dass Sinneserfahrungen und deren Verarbeitung subjektiv sind. Die oben aufgeführten Aspekte helfen, diese Erfahrungen als Gegenstand wissenschaftlicher Auseinandersetzung zu betrachten und sie wissenschaftlich zu behandeln (vgl. [3]).

Zusätzlich zu Qualitäts- und Eignungsaspekten werden die empirischen Forschungsmethoden durch folgende **Strukturaspekte** gekennzeichnet:

- **Kontrolle:** Die Ausprägung der Kontrolle kann von „kein Einfluss auf die Bedingungen, aus denen das Beobachtete entsteht“ bis „beliebige Manipulation der Bedingungen“ variieren. Bei hoher Kontrolle werden die Aspekte der Untersuchung beobachtet, an denen tatsächlich Interesse besteht (anstatt nur etwas Ähnliches) (vgl. [20]).
- **Genauigkeit:** Die Genauigkeit reicht von „direkte, objektive, vollständige und genaue Beobachtung und Messung“ bis „die Beobachtungen sind subjektive, unvollständige Eindrücke aus zweiter Hand. . . “ Von der Genauigkeit hängt die Verlässlichkeit der Resultate ab (vgl. [20]).
- **Relativität:** Die Relativität wird von „Beobachtung nur einer Sorte von Bedingungen“ bis „Vergleich verschiedener Bedingungen“ gestuft (vgl. [20]).
- **Replikation:** Die Replikation kann von „einmalige“ bis „häufige Durchführung einer beobachteten Tätigkeit“ variieren. Hohe Replikation ermöglicht den Zugriff auf mehrere Datensätze bei der Analyse der Ergebnisse und sorgt dafür, dass sich zufällige oder einmalige Effekte als solche erkennen lassen (vgl. [20]).

Beim Einsatz von empirischen Forschungsmethoden muss folgendes beachtet werden: Wird eine Hypothese mit Hilfe von empirischen Methoden als „wahr“ eingestuft, spricht das für die untersuchte Theorie. Damit wird sie aber noch nicht „bewiesen“, da die Wissenschaftler sich mittels empirischer Untersuchungen nur an die Wahrheit annähern können (vgl. [22]).

2.2 Arten von empirischen Forschungsmethoden

In diesem Abschnitt wird der Überblick über die im softwaretechnischen Bereich öfters verwendeten empirischen Forschungsmethoden gegeben. Es werden die Begriffe Fallstudie, Benchmarking, Feldstudie, Umfrage, Metastudie und kontrolliertes Experiment beschrieben.

Fallstudien dienen zur Untersuchung von einzelnen Werkzeugen, Methoden oder ihren Eigenschaften. Ihr Forschungsgegenstand wird anhand eines konkreten, meistens einfachen Beispiels untersucht. Das Beispiel kann auch komplex sein und wird in der künstlichen oder typischen Umgebung ausgeführt. Es ist möglich, mit Hilfe von Fallstudien auch mehrere Werkzeuge und Methoden zu vergleichen, da diese Forschungsmethode universell anwendbar und leicht durchzuführen ist. Der Aufwand lässt sich normalerweise an vorhandene Ressourcen und notwendige Genauigkeit anpassen. Im Gegensatz zum kontrollierten Experiment müssen nicht alle Faktoren während einer Fallstudie konstant gehalten werden. Das größte Problem bei dieser Forschungsmethode liegt bei der Interpretation: Wegen dem Einzelfallcharakter der Umgebungsbedingungen ist es selbst bei vergleichenden Fallstudien schwierig festzustellen, welche Faktoren die beobachteten Effekte hervorgerufen haben (vgl. [20]).

Standardisierte Fallstudien werden **Benchmarks** genannt. Unter Benchmark wird ein präzise definierter Anwendungsfall verstanden, durch den die Leistungsfähigkeit einer Methode ermittelt wird. Das Ziel des Benchmarks ist es, direkt vergleichbare Ergebnisse zu bekommen, wobei die Untersuchungen von unterschiedlichen Forschern ausgeführt werden können (vgl. [20]).

Feldstudien werden unter natürlichen Arbeitsbedingungen ausgeführt, um die Verallgemeinerung der in der künstlichen Umgebung erzielten Resultate zu untersuchen. Meistens sind das die Resultate von Fallstudien oder kontrollierten Experimenten. Der Hauptvorteil von Feldstudien ist die hohe Komplexität der Situationen, die sich untersuchen lassen und die Sicherheit, dass die gewonnenen Resultate zumindest auf ein reales Projekt auch wirklich zutreffen. Der Nachteil besteht darin, dass wegen der hohen Komplexität die Beschreibung der Forschung sowie die Verallgemeinerung der Ergebnisse sehr schwierig sind (vgl. [20]).

Umfragen sind relativ einfach und billig. Während einer Umfrage beantworten die Teilnehmer mehrere von den Wissenschaftlern formulierte Fragen. Diese können sich sowohl auf objektive als auch auf subjektive Sachverhalte beziehen. Die Umfrageteilnehmer gehören normalerweise zu einer speziell ausgewählten Personengruppe (Zielgruppe), deren Antworten immer subjektiv und nur schwer überprüfbar sind, daher ist die Verlässlichkeit der Ergebnisse einer Umfrage oft unklar. Die Auswertung der Resultate ermöglicht Rückschlüsse auf das subjektive Empfinden der Teilnehmer in Bezug auf den Forschungsgegenstand. Die schriftliche Befragung wird mit Hilfe von Fragebögen durchgeführt, mündliche Befragung heißt „strukturiertes Interview“ (vgl. [20]).

Der Zweck einer **Metastudie** ist die Konsolidierung des Wissens zu einem konkreten Thema aus mehreren zuvor publizierten Studien. Während einer Metastudie soll geklärt werden, ob die Resultate aus früheren Studien sich gegenseitig bestätigen oder einander widersprechen. Es wird ebenfalls untersucht, welche Aspekte aus dem gewählten Thema noch in keiner Studie beleuchtet wurden. Die Vorteile von Metastudien sind der vergleichsmäßig geringe Aufwand

und die Kosten, sowie der große Beitrag zur besseren Orientierung anderer Wissenschaftler im untersuchten Gebiet. Nachteilig ist die fehlende Möglichkeit, Lücken in vorliegenden Beobachtungen zu schließen, da während einer Metastudie kein eigenes Experiment durchgeführt wird. Die Voraussetzung für eine Metastudie ist die ausreichende Zahl der zum gewählten Thema passenden Studien, wobei sich die Studien in Anwendungsfall, Methodik oder Beschreibungsart unterscheiden können (vgl. [20]).

Da eine Vergleichstudie im Visualisierungsbereich unter kontrollierten Bedingungen durchgeführt werden sollte, wird das **kontrollierte Experiment** im nächsten Abschnitt genauer beschrieben.

2.3 Kontrolliertes Experiment

Wie der Name schon sagt, zeichnet das kontrollierte Experiment ein hoher Grad an Kontrolle aus, so dass seinen Ergebnissen besonders viel Vertrauen geschenkt wird. Im Vergleich zu den anderen empirischen Methoden werden beim kontrollierten Experiment die Bedingungen genau definiert und streng überwacht. Dadurch ergibt sich die Möglichkeit, besser zu begreifen, was die einzelnen Beobachtungen bedeuten.

Um unabhängige Variablen zu kontrollieren, werden die relevanten Aspekte der Umgebung konstant gehalten. Auf diese Weise werden höhere Genauigkeit und Reproduzierbarkeit des Experiments erreicht. Vorteilhaft beim kontrollierten Experiment ist das leichtere Verständnis, so dass die Probleme und Schwächen vom Experiment schneller aufgedeckt werden können. Außerdem lassen sich die Theorien über die Ursachen von beobachteten Effekten viel leichter aufstellen. Die Nachteile vom kontrollierten Experiment sind hoher Aufwand und vergleichsmäßig hohe Kosten (vgl. [20]). Die oben angeführte Beschreibung des kontrollierten Experiments lässt sich in folgender Definition zusammenfassen:

„Ein kontrolliertes Experiment ist eine **Studie**, bei der alle voraussichtlich für das Ergebnis relevanten Umstände ... konstant gehalten werden ..., mit Ausnahme von einem oder wenigen, die den Gegenstand der Untersuchung bilden ... Die Beobachtungen ... für verschiedene gezielt ausgesuchte Werte der Experimentvariablen ... werden miteinander verglichen, um so zu reproduzierbaren Aussagen zu kommen, die eine vor dem Experiment definierte Experimentfrage beantworten. Die Experimentfrage ist ein genügend enger Aspekt einer relevanten Forschungsfrage“ [20].

2.3.1 Zweck von kontrolliertem Experiment

Der Zweck des kontrollierten Experimentes sind die Beobachtungen, deren Ursachen eindeutig feststellbar sind: „Wenn ich etwas zweimal mache und dabei alle Umstände bis auf einen gleich sind, dann müssen eventuelle Unterschiede in den Ergebnissen E von der Änderung dieses einen Umstands U herrühren. Mit einem kontrollierten Experiment kann man also die Kausalität zwischen Ereignissen (von U zu E) untersuchen“ [20].

2.3.2 Abhängige und unabhängige Variablen

Der Begriff Variable dient der Beschreibung von Merkmalsunterschieden bei einer Gruppe von Objekten. Während des Experiments werden als Ergebnis die Werte der abhängigen Größe E gesammelt. Diese Werte hängen von dem gewählten Umstand U ab, deshalb heißt die Größe E **abhängige Variable** und der Umstand U heißt **unabhängige Variable**. Wenn die Auswirkung kombinierter Umstände untersucht werden soll, kann ein Experiment auch mehrere unabhängige Variablen enthalten. In diesem Fall muss das Experiment die Beobachtungsergebnisse für alle möglichen Kombinationen aus allen Werten dieser Variablen liefern (vgl. [20]).

Die Menge aller Merkmalsmessungen wird als **quantitative** Daten bezeichnet, falls Merkmale oder Merkmalsausprägungen verbal beschrieben werden, so heißen die Daten **qualitativ**. Bei quantitativen Variablen sind **stetige** (kontinuierliche) und **diskrete** (diskontinuierliche) Variablen zu unterscheiden. Die Eigenschaft von einer stetigen Variablen besteht darin, dass sie in jedem beliebigen Intervall unendlich viele Merkmalsausprägungen besitzen kann (z. B. die Variablen Länge oder Zeit). Eine diskrete Variable besitzt in einem begrenzten Intervall endlich viele Ausprägungen (z.B. Anzahl von jährlichen Geburten) (vgl. [3]).

Für unterschiedliche Typen von Variablen existieren unterschiedliche Arten von Skalen. Die **Nominalskala** teilt Objekte in Klassen ein wie z. B. “männlich” oder “weiblich”, ohne irgendeine Gewichtung vorzunehmen. Die Klassenbildung muss vollständig sein, d. h. die Kategorien erfassen alle möglichen Fälle. Gleichzeitig muss eine eindeutige Zuordnung erfolgen (gegenseitiger Ausschluss), so dass ein Objekt nur zu einer Kategorie gehören kann. Oft lassen sich Variableneigenschaften nach ihrem Ausprägungsgrad ordnen, auch wenn das nur unpräzise erfolgen kann. Die Skala für diese Variablen heißt **Rang-** oder **Ordinalskala.** Ein Beispiel für eine unpräzise Unterscheidung der Eigenschaften ist das dreigliedrige Schulsystem: Die Lerninhalte auf dem Gymnasium haben ein höheres Anspruchsniveau als die auf der Realschule, und auf der Hauptschule ist das Niveau am niedrigsten. Eine **Intervallskala** besitzt eine definierte Maßeinheit, so dass die Objekte wie bei einer Rangskala nach der Stärke der Merkmalsausprägung geordnet werden können. Außerdem wird diese Merkmalsausprägung durch konkrete Zahlen ausgedrückt. Diese Zahlen reflektieren gleiche Differenzen in der Ausprägung des Merkmals. Ein Beispiel für die Verwendung einer Intervallskala ist die Punktvergabe in einem Intelligenztest. Im Unterschied zur Intervallskala existiert bei der **Verhältnisskala** ein absoluter Nullpunkt. Gewicht, Abstand und Temperatur in Kelvin werden auf einer Nominalskala gemessen (vgl. [14]).

2.3.3 Zu kontrollierende Variablen

Es existiert noch eine dritte Art von Variablen, nämlich die **zu kontrollierenden Variablen** oder **Störvariablen**, deren wichtigste Ursache die individuellen Unterschiede von Probanden sind. Anders ausgedrückt, Störvariablen sind „alle Einflussgrößen auf die abhängige Variable, die in einer Untersuchung nicht erfasst werden...“ [3].

Beim kontrollierten Experiment müssen alle Umstände außer den absichtlich manipulierten

gleich gehalten werden. Das ist in der realen Untersuchung so gut wie unmöglich: Der bestimmte Zustand der Welt lässt sich nicht zu 100 Prozent wiederholen. Außerdem ist es nicht leicht, alle relevanten Störvariablen überhaupt zu berücksichtigen. Zum Glück muss dies auch nicht angestrebt werden: Fast alle relevanten Störvariablen, ob bekannt oder nicht, lassen sich mit derselben Methode kontrollieren. Diese Methode besteht aus **Replikation** und **Randomisierung**. Replikation wird dadurch erreicht, dass nicht das Verhalten einer einzigen Person für jeden Wert der unabhängigen Variablen betrachtet wird, sondern das Verhalten einer ganzen Gruppe. Da die Versuchspersonen normalerweise zufällig gewählt werden, ist auch Randomisierung gewährleistet (vgl. [20]). Um die Replikation und Randomisierung zu ermöglichen, soll die Gruppe von Versuchspersonen vergleichsmäßig groß sein. Empfohlen wird eine Anzahl von 30 bis 50 Teilnehmern (vgl. [28]). So gleichen sich die individuellen Unterschiede der Probanden aus und die Kosten des Experiments halten sich in Grenzen.

Um ein maximal fehlehrfreies und fundiertes Experiment zu bekommen, ist es wichtig, die im Abschnitt 2.4 aufgezählten möglichen Schwierigkeiten zu überwinden.

2.4 Innere Gültigkeit von einem kontrollierten Experiment

Unter dem Begriff „innere Gültigkeit“ wird der Grad der Kontrolle über Störvariablen verstanden. Es existieren folgende Bedrohungen für innere Gültigkeit:

- Reifung
- Instrumentation
- Historie
- Auswahl
- Regression
- Sterblichkeit
- Anforderungscharakteristik
- Verarbeitungsfehler

Diese Faktoren können die Ergebnisse eines Experiments verzerren und müssen mittels speziell entwickelter Maßnahmen kontrolliert werden (vgl. [20]).

Reifung: Unter diesem Begriff werden die Veränderungen im Verhalten von Versuchspersonen verstanden, die über den längeren Zeitraum hinweg auftreten. **Lern-** und **Reihenfolgeeffekte** treten vor allem dann in Erscheinung, wenn mehrere Aufgaben hintereinander gelöst werden müssen. Normalerweise lässt bei späteren Aufgaben die Konzentration nach, weil eine **Ermüdung** eintritt. Es besteht aber auch die Möglichkeit, dass die Leistung besser wird oder eine Änderung der Arbeitsweise eintritt. Diese Effekte sind darauf zurückzuführen, dass die Versuchsperson aus früheren Aufgaben lernt, mit einem Typ von Aufgaben “besser zu Recht zu

kommen oder weil sich inhaltliche Erkenntnisse aus der Bearbeitung einer Aufgabe profitabel zur Bearbeitung einer späteren einsetzen lassen“ [20]. Es sind aber auch negative Reihenfolgeeffekte möglich, da die Versuchsperson durch den Inhalt der vorherigen Aufgabe verwirrt sein kann. Bei der Erstellung eines Experimententwurfs müssen diese Faktoren beachtet und Alles dafür unternommen werden, dass die Ermüdung sowie Lern,- und Reihenfolgeeffekte keinen Einfluss auf die Messung von abhängigen Variablen haben (vgl. [20]).

Instrumentation: Die Verfälschung der Messung von abhängigen Variablen ist auch dann wahrscheinlich, wenn sich das Verhalten vom Experimentator über die Zeit ändert. Seine absichtlichen oder unabsichtlichen Bemerkungen und Kommentare während des Experimentablaufs haben Einfluss auf das Wohlbefinden und die Motivation vom Probanden. Daher ist es wichtig, dass sich der Experimentator konstant verhält und der Stärke seines Einflusses bewusst ist. Eine weitere Ursache für eine fehlerhafte Messung ist ein falscher Experimentaufbau: Ein Programmwerkzeug für die automatisierte Messung könnte beispielsweise bei späteren Versuchsdurchführungen wegen der zunehmenden Dateifragmentierung der Festplatte langsamer werden (vgl. [20]).

Historie: Das Phänomen tritt bei Experimenten auf, die über mehrere Wochen laufen. Die Zeit, die zwischen zwei Versuchen liegt, kann auf den Zustand und die Motivation der Teilnehmer positiv oder negativ wirken. Es ist auch möglich, dass ehemalige Teilnehmer den künftigen Teilnehmern die Information über Versuchsablauf, Aufgaben oder Lösungen verraten (vgl. [20]).

Auswahl: Manchmal kann die Bildung von Versuchsgruppen nicht wirklich zufällig erfolgen. In einigen Fällen werden für unterschiedliche Gruppen verschiedene Vorkenntnisse notwendig. Die entsprechende Ausbildung ist im Rahmen eines Experiments wegen des hohen Aufwands meistens nicht möglich. Die Auswahleffekte bedrohen in diesem Fall die Gültigkeit des Experiments, da es ungewiss ist, ob sich die Kriterien für die Gruppenbildung auf die Ergebnisse auswirken oder nicht (vgl. [20]).

Regression: Regression zum Mittelwert ist ein Spezialfall der Auswahl. Dieses Phänomen kann eintreten, wenn eine Versuchsperson beim ersten Versuch eine für ihre Verhältnisse besonders gute (oder besonders schlechte) Leistung gezeigt hat. Dann ist die Wahrscheinlichkeit groß, dass bei der nächsten Messung die Leistung derselben Versuchsperson schlechter (bzw. besser) sein wird. Die Gültigkeit eines Experiments wird durch eine nichtzufällige Einteilung in die neuen Versuchsgruppen, von denen eine „gut“ und eine „schlecht“ ist, gefährdet (vgl. [20]).

Sterblichkeit: Es besteht oft die Möglichkeit, dass die Teilnehmer vom Experiment ausscheiden. Erfolgt dies aufgrund von Frustration, ist die Gültigkeit des Experiments wieder verletzt, da nur die tendenziell leistungsfähigeren Probanden übrigbleiben. Der Einfluss von Sterblichkeit ist dann besonders hoch, wenn die Teilnehmer in Gruppen eingeteilt sind und aus einer Gruppe mehr Probanden ausscheiden als aus den anderen (vgl. [20])

Anforderungscharakteristik: Es ist nicht selten, dass die Experimentatoren in Bezug auf manche Versuchsvariablen nicht neutral eingestellt sind, sondern sie erhoffen sich von bestimmten Probanden eine bessere Leistung als von den anderen. Die Versuchsleiter könnten diese Probanden unbewusst freundlicher behandeln, was logischerweise Auswirkungen auf die Motivati-

on haben kann. In solch einem Fall wird die Motivation zu einer relevanten und nicht kontrollierten Störvariablen (vgl. [20]).

Verarbeitungsfehler: Dieser Fehlertyp kann in vielen unterschiedlichen Fällen auftreten. Beispielsweise kann die abhängige Variable falsch gemessen oder falsch verarbeitet werden. Die Ursache dafür könnte eine unzuverlässig arbeitende automatisierte Messvorrichtung oder ein einfacher Tippfehler beim Eingeben von Datenwerten sein. Ein weiteres Problem könnte eine unsinnige oder falsch angewandte statistische Methode der Datenverarbeitung sein, die offensichtlich falsche Ergebnisse liefern wird (vgl. [20]).

Insgesamt existieren hunderte möglicher Bedrohungen für innere Gültigkeit. Die meisten lassen sich jedenfalls in die oben beschriebenen Kategorien einordnen. Die wichtigste Aufgabe der Experimentdesigner besteht darin, jede dieser Kategorien auf die relevanten Gefahren zu untersuchen und Maßnahmen zum Vermeiden dieser Gefahren zu entwickeln (vgl. [20]).

2.5 Wann ist ein Experiment gut?

Ein Experiment muss zwei Eigenschaften besitzen, damit es als guter Forschungsbeitrag bezeichnet werden kann: Es muss relevant und glaubwürdig sein.

Die **Relevanz** des Experiments besteht aus zwei Aspekten. Der erste Aspekt ist die Wichtigkeit der Forschungsfrage, die im Experiment untersucht wird. Der zweite Aspekt ist der Beitrag, den die Ergebnisse des Experiments zur Beantwortung dieser Forschungsfrage liefern. Von der Relevanz eines Experiments hängt oft ab, ob das Experiment von den potentiellen Lesern überhaupt wahrgenommen wird (vgl. [20]).

Die **Glaubwürdigkeit** besteht aus drei Komponenten: innerer Gültigkeit, äußerer Gültigkeit und Schlussfolgerungen. Die innere Gültigkeit wurde im Abschnitt 2.4 ausführlich beschrieben. „Die äußere Gültigkeit eines kontrollierten Experiments ist der Grad, in dem sich seine Resultate korrekt auf andere Anwendungsfälle übertragen lassen – insbesondere auf solche, die in der Praxis häufig vorkommen“ [20]. Das bedeutet, die Resultate müssen hinreichend verallgemeinerbar sein und außerdem muss vermittelt werden, wo diese Verallgemeinerbarkeit ihre Grenzen hat. Als dritte Komponente der Glaubwürdigkeit gelten die guten Schlussfolgerungen. Sie dürfen einerseits nicht zu weitreichend, andererseits nicht zu vorsichtig sein. Wichtig ist, dass in den Schlussfolgerungen die wissenschaftlichen und praktischen Konsequenzen „prägnant aufgezeichnet werden, dabei aber angemessen bleiben“ [20].

2.6 Vergleich der Methoden

In Tabelle 1 sind die spezifischen Stärken und Schwächen jeder Methode zusammengefasst.

Aspekt	Fallstudie	Feldstudie	kontrolliertes Experiment	Umfrage	Metastudie
Aufwand der Forscher	oOO	oOO	OO	oO	oO
Aufwand der Teilnehmer	oOO	o	oOO	o	-
Verlässlichkeit der Resultate	O	oO	O	o	o
Beschreibbarkeit	OO	oO	OO	O	O
Wiederholbarkeit der Forschung	oOO	o	O	O	O
Reproduzierbarkeit der Resultate	oO	oO	OO	oO	O
Verallgemeinerbarkeit der Resultate	oOO	oO	oOO	oOO	oOO
Genauigkeit	oO	oO	O	o	oOO
Vergleich	oO	oOO	O	OO	OO
Replikation	o	oOO	O	O	OO

Tabelle 1: Typische Eigenschaften der verschiedenen Forschungsmethoden [20].

Obwohl die benutzte Skala nur drei Stufen enthält, lassen sich die Eigenschaften einer Methode keiner Stufe eindeutig zuordnen. Das liegt daran, dass die Ausprägung jeder Eigenschaft vom einzelnen konkreten Fall abhängt (Ausprägung: o - gering, o - mittel, O - hoch, - - entfällt) (vgl. [20]).

2.7 Ein Experiment unter Einsatz des Eye-Trackers

Die Augen sind die wichtigsten Sinnesorgane des Menschen in Bezug auf die übertragene Datenmenge. Sehende Menschen nehmen 80% der Information visuell war. In der folgenden Liste sind die Mengen der Daten aufgeführt, die die wichtigsten Sinnesorgane in einer Sekunde übertragen:

- Augen 10 Mio. Bit/sec
- Haut 1 Mio. Bit/sec
- Ohren 100 000 Bit/sec
- Nase 100 000 Bit/sec (vgl. [8, 9]).

Diese Werte bilden einen wichtigen Hinweis darauf, dass die Beobachtung des menschlichen Sehens eine große Rolle in einer empirischen Untersuchung spielen kann. Das gilt sowohl für die Untersuchung psychologischer Probleme als auch für die Erforschung der Benutzbarkeit. Dank moderner Eye-Tracker-Geräte ist eine präzise Aufzeichnung der Augenbewegungen möglich geworden. Unter Verwendung eines solchen Instruments wurde im Rahmen dieser Diplomarbeit eine Vergleichsstudie durchgeführt. Das Ziel der Studie bestand darin, den Einfluss von neun unterschiedlichen Anordnungen der Knoten-Kanten-Diagramme auf die Lösbarkeit von Visualisierungsaufgaben zu untersuchen. Die Blickbewegungen der Probanden wurden mit Hilfe eines Eye-Tracker-Gerätes verfolgt und aufgezeichnet. In den Abschnitten 2.7.1 bis 2.7.3 folgt ein Überblick über die Geschichte sowie über Vor- und Nachteile des Eye-Trackings.

2.7.1 Geschichte des Eye-Trackings

„Unter dem Eye-Tracking versteht man das Aufzeichnen der hauptsächlich aus Fixationen (Punkte, die man genau betrachtet) und Sakkaden (schnellen Augenbewegungen) bestehenden Blickbewegungen einer Person“ (vgl. [9]).

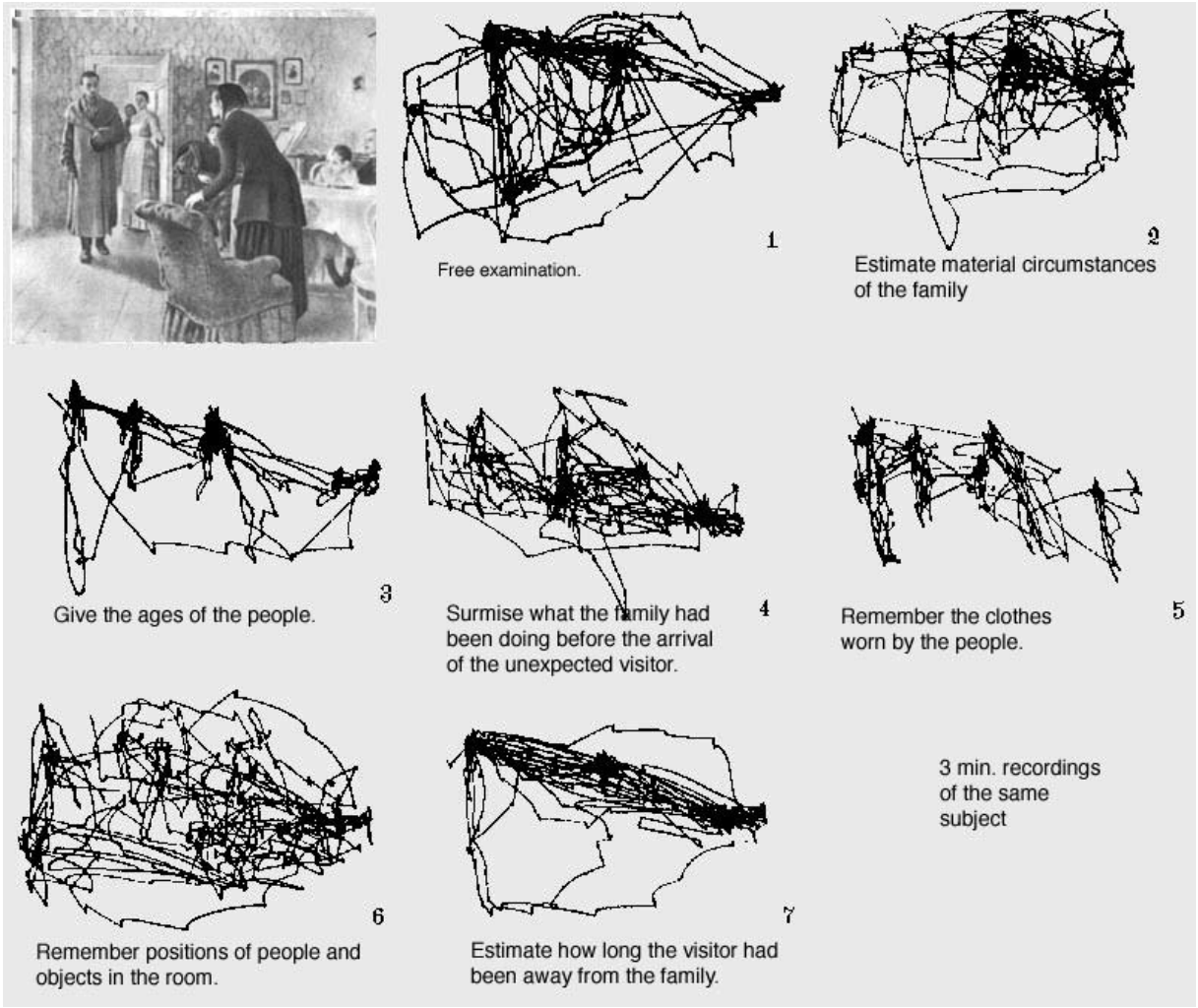

Abbildung 3: Eye-Tracking des Bildes “The Visitor” aus den Untersuchungen von Yarbus [32]

Eye-Tracking-ähnliche Untersuchungen sind schon seit langem bekannt: Schon im 11. Jahrhundert erforschte ein ägyptischer Arzt die Augenbewegungen und beschrieb sie als Folge schneller Einzelbewegungen. Im 19. Jh. beschrieb der Franzose Emile Java als einer der ersten die Augenbewegungen beim Lesen. 1908 entwickelte Huey eine Messeinrichtung zur direkten Messung der Augenbewegung. Die Messeinrichtung bestand aus einer keramischen Haftschale mit einem Loch in der Mitte und einem Aluminiumzeiger. Die Haftschale wurde direkt auf die Hornhaut des Auges aufgebracht. Der Aluminiumzeiger zeichnete die Bewegungen des Augapfels auf einem Papierstreifen nach (vgl. [9, 24]).

Am Anfang des 20. Jahrhunderts wurde eine Filmkamera erfunden. Die Aufnahmen der Kameras gaben die Möglichkeit, die Augenbewegungen aufzuzeichnen und nachträglich zu analysieren. Judd und Buswell entwickelten unter Einsatz einer Kamera ein Messverfahren, welches

den direkten Kontakt von Geräten mit dem Auge vermied. Seitdem wird Eye-Tracking systematisch im psychologischen Bereich verwendet, wobei der Schwerpunkt auf der Aufzeichnung der Augenbewegungen beim Lesen liegt.

In den 50er Jahren entdeckte der russische Psychologe Alfred Lukjanovic Yarbus, dass Menschen ein Bild je nach Aufgabenstellung unterschiedlich betrachten (Abb. 3). 1967 brachte er sein Buch „Eye Movements and Vision“ heraus. Das Buch hatte damals und hat heute noch einen großen Einfluss auf die Entwicklung des Eye-Trackings.

In den 1980er Jahren stellten Just und Carpenter eine Eye-Mind-Hypothese auf, die besagt, dass es keine bedeutende Zeitverzögerung zwischen dem Fixieren eines Punktes und der Verarbeitung dieses Eindrucks gibt. In demselben Zeitraum begannen Wissenschaftler mit Hilfe von Eye-Tracking, die Fragen der Mensch-Computer-Interaktion zu untersuchen. Da in den letzten Jahren die technologischen Entwicklungen wie Internet, E-Mail und Videokonferenzen zu unentbehrlichen Mitteln des Informationsaustauschs geworden sind, setzen Forscher das Eye-Tracking immer öfter zur Untersuchung der Benutzbarkeitsfrage ein (vgl. [9, 24]).

2.7.2 Vorteile

Einer der wertvollsten Aspekte des Eye-Trackings ist die Möglichkeit festzustellen, ob die angezeigten Stimuli tatsächlich angesehen werden. Die weiteren Vorteile sind die folgenden:

- Die Blickbewegungen geben Hinweise darüber, was ein Benutzer betrachtet oder liest, wie lange und in welcher Reihenfolge.
- Durch eine Analyse der Dauer und Anzahl von Fixationen und Sakkaden lässt sich klären, ob ein Nutzer sich auf den Inhalt konzentriert oder ihn nur oberflächlich betrachtet.
- Mit Hilfe von Eye-Tracking kann ermittelt werden, welche Bereiche des Bildschirms die besondere Aufmerksamkeit erhalten.
- Bei neuer Information lässt sich anhand von Veränderungen des Pupillendurchmessers erkennen, ob unbekannte, irrelevante oder erwartete Begriffe und Bereiche betrachtet werden.
- Mittels der Eye-Tracking-Daten lassen sich Strategien untersuchen und vergleichen, die verschiedene Nutzer bei der Lösung der gestellten Aufgaben entwickeln.
- Das Verfahren macht es möglich, Unsicherheiten und Verzögerungen in der Reaktion der Nutzer zu entdecken, die mit anderen Methoden nicht zuverlässig messbar sind. Dadurch lassen sich menschliches Verhalten und häufig auch seine Denkweise besser verstehen.
- Mit Hilfe von Eye-Tracking lassen sich komplizierte psychologische Prozesse erforschen, über die ein Mensch von sich aus schlecht berichten kann.

- Die Eye-Tracking-Methode ist oft zuverlässiger als die Methoden, bei denen die Beschreibung des eigenen Verhaltens möglich bzw. notwendig ist. Die Gefahr bei diesen Methoden besteht darin, dass die Beschreibung des eigenen Verhaltens aufgrund von sozialen Erwartungen verfälscht werden könnte, um so einen besseren Eindruck zu hinterlassen.

Visuelle Suche, Szenenwahrnehmung, auditive Verarbeitung der Sprache, Problemlösen, Mensch-Computer Interaktion, Sport, Luftfahrt und das Lesen sind nur ein Teil der Fragen, die mit Eye-Tracking untersucht werden können (vgl. [8, 9, 10, 24, 33]).

2.7.3 Nachteile

Neben den oben erwähnten Vorteilen hat Eye-Tracking wie jede andere Methode auch ihre Nachteile.

- **Kosten**: Die Leistungsfähigkeit und Flexibilität der modernen Eye-Tracker-Geräte implizieren deren hohen Preis. Weitere Kosten entstehen durch die Bezahlung der Studienteilnehmer und deren mögliche Ausscheidung aus einem Experiment (vgl. [26]).
- **Aufwand:** Die Forschung mit Hilfe eines Eye-Tracker-Gerätes sollte im Rahmen eines kontrollierten Experiments durchgeführt werden, deren gründliche Vorbereitung und fehlerfreie Durchführung normalerweise sehr aufwendig ist.
- **Einfluss von Qualifikation:** Bei bestimmten Untersuchungen vor allem im Visualisierungsbereich können Alter, Qualifikation oder Erfahrung einen großen Einfluss auf die Ergebnisse der Studie haben und diese verfälschen (siehe auch Begriff „Auswahl“ im Abschnitt 2.4)
- **Komplexität von kognitiven Prozessen:** Es passiert nicht selten, dass der Nutzer zwar die Dinge mit dem Blick fixiert, aber seine Gedanken nicht um das gerade Gesehene kreisen (vgl. [9]). In diesem Fall sind keine zuverlässigen Schlussfolgerungen über die wahrscheinlichen Lösungsstrategien möglich bzw. selbst die Trennung von „brauchbaren“ und „nicht brauchbaren“ Blickfixierungen praktisch unmöglich.
- **Peripheres Sehen:** Sehr viel Information gelangt durch die Peripherie des Sehfeldes in das kognitive System des Menschen (vgl. [9]). Diese Information bleibt aber für das Eye-Tracking unerschlossen, da das Eye-Tracking nur foveales Sehen verfolgt, bei dem das Auge exakt auf einen Punkt gerichtet sein muss (vgl. [31]).

Trotz der aufgelisteten Nachteile ist es offensichtlich, dass das Eye-Tracking-Verfahren eine für die Untersuchung der Visualisierungstechniken sehr gut geeignete Methode ist.

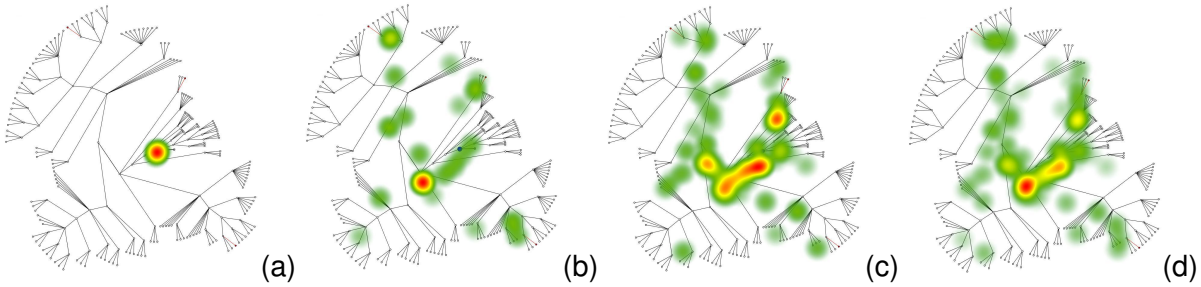

Abbildung 4: Heat Maps von drei einzelnen Betrachtern (a)-(c); Zusammenfassung der drei Heat Maps (d)

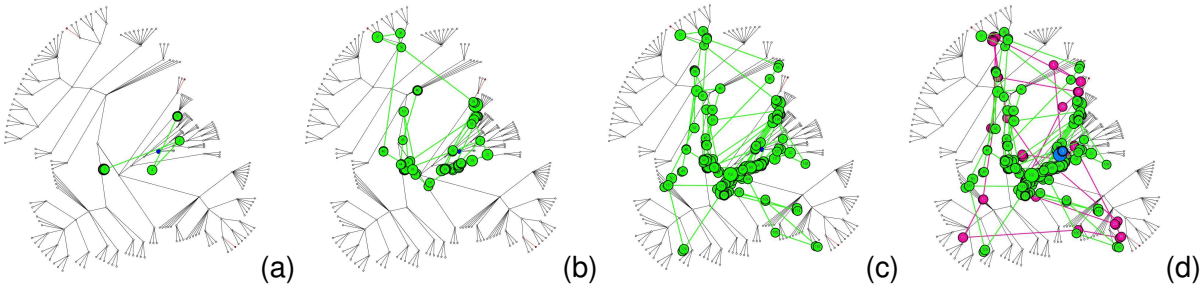

Abbildung 5: Gaze Plots eines Betrachters rück- und vorgespult (a)-(c); Gaze Plots mehrerer Betrachter (d)

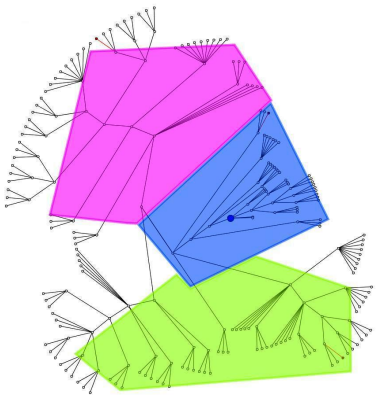

Abbildung 6: Benutzerdefinierte AOIs: blau - AOI_1, rot - AOI_2, grün - AOI_3

2.7.4 Visualisierung der Eye-Tracker-Ergebnisse

Die Daten, die ein Eye-Tracking-System liefert, werden normalerweise in folgenden drei Formen präsentiert:

- Rohdaten
- graphische Darstellungen
- Darstellung der Ergebnisse statistischer Analyse

Die Bestandteile der Rohdaten sind die Fixationskoordinaten, die Fixationsdauer, die Sakkadenwinkel und oft auch der Pupillendurchmesser. Diese Rohdaten können statistisch analysiert und graphisch dargestellt werden (vgl. [9]).

Die drei sehr verbreiteten und für die Untersuchung der Lösungsstrategien der Probanden besonders gut geeigneten graphischen Darstellungsarten sind Heat Maps, Gaze Plots und Gaze-Replays (Abb. 4, Abb. 5).

Die Heat Maps (Gaze Spots/Hotspots) sind Bereiche, die von den getesteten Probanden besonders oft und intensiv betrachtet wurden. Mit Heat Maps lässt sich das Blickverhalten von einer ganzen Gruppe von Testpersonen darstellen (Abb. 4) (vgl. [9]). Die Heat Maps können abhängig von der Anzahl oder von der Dauer der Fixationen dargestellt werden.

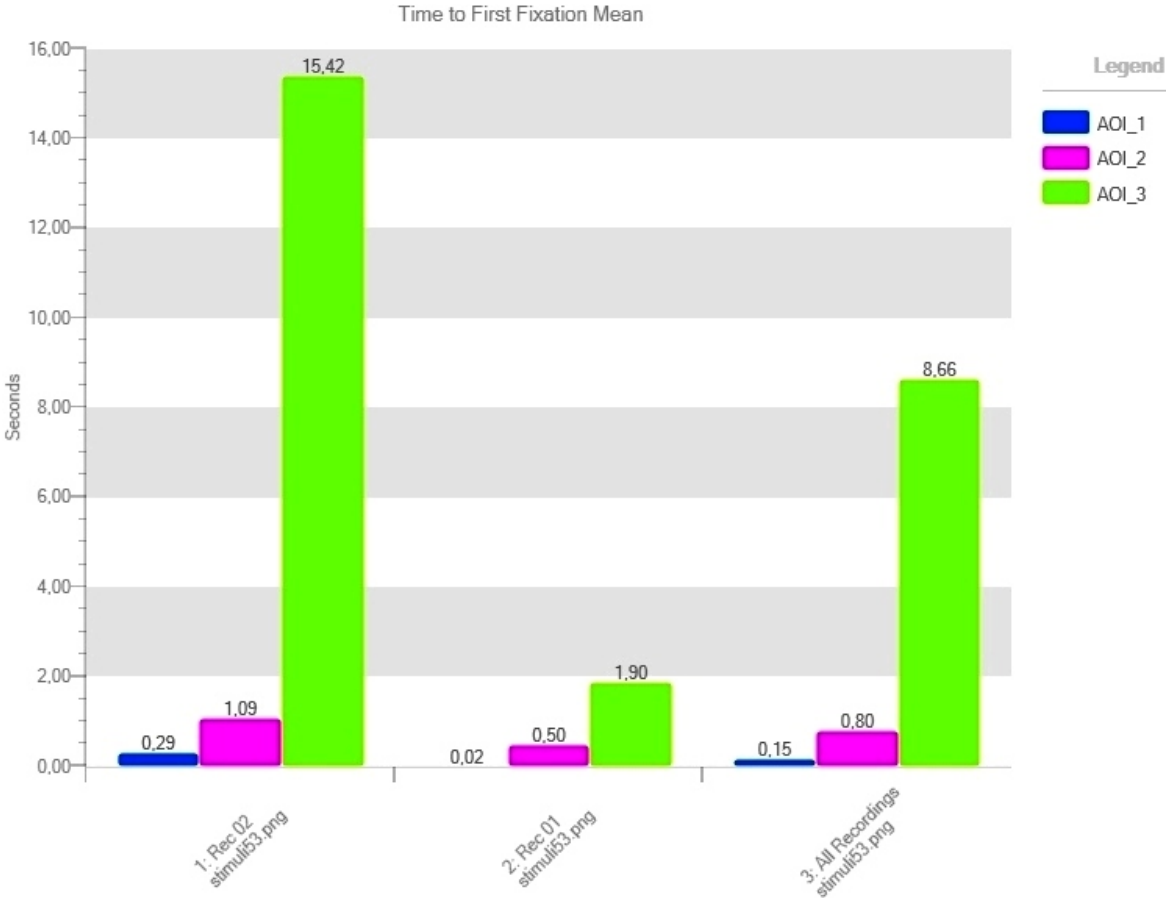

Abbildung 7: Darstellung der Ergebnisse der statistischen Analyse in Diagrammform

Die Gaze Plots (Gaze-Traces-Darstellung) stellen die Blickbewegungen eines einzelnen Probanden als Sakkaden (Linien) und Fixationen (Kreise) dar (vgl. [9]). Die Fixationen sind der Reihenfolge nach nummeriert und ihre Größe entspricht der Fixationsdauer. Das in dieser Diplomarbeit verwendete Eye-Tracker-System besitzt eine manuelle Vor- und Rückspulfunktion, so dass jede einzelne Fixierung schrittweise betrachtet und analysiert werden kann (Abb. 5).

Die Gaze-Replays sind Videos der Blickpfade (Gaze-Traces) von Probanden, die in Echtzeit oder Zeitlupe abgespielt werden können (vgl. [9]). Es existiert ebenfalls die Möglichkeit, die Gaze-Replays manuell vor- und zurückzuspulen. Dabei lässt sich sogar die Veränderung der Größe der einzelnen Fixationen beobachten, die die Dauer der Fixierung wiederspiegelt.

Der in dieser Diplomarbeit verwendete Eye-Tracker enthält statistische Werkzeuge. Sie basieren auf den AOIs (Areas of Interest). Für die statistische Analyse müssen diese zuerst definiert

	Time to First Fixation								
	stimuli53.png								
	AOI_1			AOI_2			AOI_3		
Recordings	N (Count)	Mean (Seconds)	Sum (Seconds)	N (Count)	Mean (Seconds)	Sum (Seconds)	N (Count)	Mean (Seconds)	Sum (Seconds)
Rec 02	1	0,29	0,29	1	1,09	1,09	1	15,42	15,42
Rec 01	1	0,02	0,02	1	0,50	0,50	1	1,90	1,90
All Recordings	2	0,15	0,31	2	0,80	1,59	2	8,66	17,32

Tabelle 2: Darstellung der Ergebnisse der statistischen Analyse in Tabellenform

werden (Abb. 6). Die Ergebnisse der Analyse werden in Diagramm- und Tabellenform dargestellt (Abb. 7, Tab. 2). Bei der Analyse der AOIs besteht die Möglichkeit, verschiedene Metriken zu wählen: “Time to First Fixation”, “Fixations Before”, “First Fixation Duration”, “Total Fixation Duration”, “Fixation Count” usw. Das Statistiktool enthält die Funktion zum Exportieren der Daten sowie viele weitere Einstellungs- und Auswahlmöglichkeiten.

3 Fragen zur Durchführung einer Studie

Ein grundlegendes Ziel der Durchführung von Vergleichsstudien im Visualisierungsbereich ist der Versuch zu verstehen, warum und wann eine bestimmte Technik gut und wann schlecht funktioniert. Es ist wichtig festzustellen, für welche Arten von Aufgaben und unter welchen Bedingungen eine bestimmte Methode qualitativ hochwertige Ergebnisse liefert (vgl. [15]).

Das Ausführen einer Studie erfordert erhebliche Ressourcen. Um zuverlässige Schlussfolgerungen erzielen zu können, sind ein klares Ziel, kontrollierte Laborbedingungen sowie genaue und begrenzte Aufgaben notwendig (vgl. [28]). Diese Anforderungen sind erfüllt, wenn eine Vergleichsstudie in Form eines kontrollierten Experiments durchgeführt wird. Die folgende Beschreibung von sieben Phasen gilt für kontrollierte Experimente und für alle Studien, die diese Form annehmen sollen.

3.1 Sieben Phasen einer Studie

In diesem Abschnitt werden die grundlegenden Aspekte aufgelistet, die vor, während und nach einer Studie beachtet werden sollten. Diese Aspekte lassen sich entsprechend der Phase der Studie gruppieren. Dabei können die Phasen in der Praxis “selten streng hintereinander ausgeführt werden, sondern wechseln einander in einem iterativen Prozess ab, der von Fall zu Fall verschieden aussieht” [20].

Es werden die folgenden Phasen vorgestellt:

- Phase der Anforderungsbestimmung
- Entwurfsphase
- Phase der Implementierung
- Phase des Tests
- Ausführungsphase
- Phase der Auswertung
- Phase der Publikation

Die im besonderen Maße für den Vergleich von Visualisierungstechniken passenden Aspekte sind kursiv hervorgehoben. Die meisten zu klärenden Fragen wurden von Prechelt [20] vorgeschlagen, die anderen Quellen werden explizit angegeben.

Die erste Phase in der Durchführung einer Studie ist die Phase der **Anforderungsbestimmung.** Hier findet die Auswahl einer Fragestellung statt, die dann als Experimentfrage untersucht werden soll. In dieser Phase sind folgende Aspekte zu klären:

- Wie lauten die Forschungsfragen, zu deren Beantwortung das Experiment einen Beitrag liefern soll? Warum sind sie relevant?

- Welche konkreten Fragen lassen sich realistischerweise im Experiment beantworten? Prüfung: Sind sie relevant?

- Unter welchen Randbedingungen wird das Experiment stattfinden müssen? Prüfung: Sind diese akzeptabel [20]? Die Frage nach Randbedingungen lässt sich genauer ausformulieren, indem sie in zwei Teilfragen aufgeteilt wird: Welche Ressourcen sind vorhanden? Welche Begrenzungen gibt es? Für die Beantwortung dieser Fragen sollen finanzielle Mittel, Zeitgrenzen, Räume, Geräte und Anzahl von Versuchsleitern in Betracht gezogen werden (vgl. [23]).

- *Was wollen wir konkret vergleichen* (vgl. [23])*? Wie in der Einführung erwähnt wurde, enthalten viele moderne Anwendungen einige Visualisierungskomponenten. Es besteht ein großer Unterschied, ob die Visualisierungstechniken im medizinischen, technischen oder kaufmännischen Bereich untersucht werden müssen.*

Wenn in der Phase der Auswertung für die untersuchten Forschungsfragen keine signifikanten Unterschiede festgestellt wurden, könnten die Wissenschaftler nachträglich neue Fragestellungen formulieren und nach weiteren signifikanten Unterschieden in Subgruppen suchen wollen. So wird die Möglichkeit gebildet, dem interessierten Leser der Vollständigkeit halber prägnante Ergebnisse zu präsentieren (vgl. [25]). Dabei heißen die Unterschiede zwischen Messgrößen dann signifikant, wenn die Wahrscheinlichkeit, dass sie zufällig so zustande gekommen sind, sehr niedrig ist. Wenn **Signifikanz** nachgewiesen wurde, kann statistisch darauf geschlossen werden, dass tatsächlich ein Unterschied zwischen erhobenen Messwerten vorliegt. Dennoch können auch Unterschiede, die statistisch signifikant sind, zufällig entstehen (vgl. [2]).

Bei der nachträglichen Suche nach signifikanten Unterschieden in Subgruppen tritt das Problem der multiplen Signifikanztests auf. Dieses Problem besteht darin, dass die vorgegebene Grenze für die Wahrscheinlichkeit fälschlicherweise auf einen signifikanten Unterschied zu schließen, nicht mehr eingehalten wird. Üblicherweise liegt diese Grenze bei 5% der Wahrscheinlichkeit. Tabelle 2 zeigt die Wahrscheinlichkeit für mindestens ein falsch positives Resultat in der Situation, dass mehrere zum 5%-Niveau ausgeführte Tests unabhängig, d.h. in sich nicht überlappenden Subgruppen durchgeführt wurden (vgl. [25]).

Anzahl unabhängiger statistischer Tests	Wahrscheinlichkeit für mindestens eine falsch positive Entscheidung
1	0.05
2	0.10
3	0.14
4	0.19
5	0.23
10	0.40
50	0.92
100	0.99

Tabelle 3: Wahrscheinlichkeit für mindestens eine falsch positive Entscheidung in Abhängigkeit von der Anzahl der durchgeführten unabhängigen statistischen Tests [25]

Die Ursache für solch irreführende Resultate liegt in einer ausgedehnten nicht von vornherein adäquat geplanten Subgruppenanalyse (vgl. [25]).

Aufgrund des Zusammenhangs zwischen der Anzahl der Tests und der Wahrscheinlichkeit für eine falsch positive Entscheidung soll aber kein Endruck entstehen, es wäre grundsätzlich ein Fehler, jemals Untersuchungen in bestimmten Subgruppen durchzuführen. Es ist ganz selbstverständlich, dass bei der Ausführung einer groß angelegten, aufwendigen und teuren Studie nach zusätzlichen Aspekten geforscht werden kann. Wichtig ist dabei, dass die Studie gemäß dieser Zielsetzung schon den ersten Phasen entsprechend geplant und durchgeführt wird (vgl. [25]).

Nach der Phase der Anforderungsbestimmung folgt die **Entwurfsphase**. Hier soll die Grobstruktur des Experiments festgelegt werden. Viele Aspekte können aufgrund von Lern- und Reihenfolgeneffekten erst in der Implementierungsphase bestimmt werden, da sie von den konkreten Aufgabenstellungen abhängig sind. Für den Entwurf sind folgende Fragen relevant:

- Welche konkreten Aufgaben sind zum Erreichen des Ziels am Besten geeignet (vgl. [23])?
- Welche unabhängigen Variablen sollen manipuliert werden? Wie viele und welche Werte sollen sie annehmen?

  Es sollten nicht mehr als drei unabhängige Variablen gleichzeitig variiert werden, andernfalls werden sowohl die Ergebnisse unübersichtlich als auch die Analyse aufwändig.
- Welche abhängige Variablen müssen/sollen/können gemessen werden?
- Welches Skalenniveau ist mit dem Typ der abhängigen Variablen verbunden (vgl. [3])?
- Wie erfolgt das Messen der abhängigen Variablen? (technische Infrastruktur)
- Gewährleistet die Messart, dass das, was gemessen werden soll, auch tatsächlich gemessen wird (vgl. [3])?
- Kann die Messung hinreichend zuverlässig und genau erfolgen?
- In welchem Ausmaß ist mit “Messfehlern” zu rechnen ([3])?
- *Ist der Einsatz von Testdaten notwendig? Zur Reproduzierbarkeit und Überprüfbarkeit müssen Testdaten klar definiert, spezifiziert und kontrolliert werden. Darüber hinaus muss ein Satz von Standarddaten vorhanden sein, der für die Visualisierungsgemeinschaft verfügbar sein sollte* (vgl. [23])*. Wenn die Frage positiv beantwortet wird, bleibt zu klären,*
- *ob synthetische oder reale Daten eingesetzt werden müssen. Letztendlich sollte die Auswertung einer Visualisierungstechnik mit Echtdaten durchgeführt werden. Aber es ist sinnvoll, mit den voll spezifizierten und systematisch kontrollierten Datenstrukturen, die in die synthetischen Daten eingearbeitet werden, zu beginnen. Die Verwendung von computergenerierten Daten ist flexibel und erlaubt ein einfacheres Entdecken von Fehlern und Ungenauigkeiten in der Visualisierungstechnik als mit den Echtdaten* (vgl. [23]).

- Werden qualitative oder quantitative Daten erhoben? Falls quantitative Daten erhoben werden, sind abhängige Variablen stetig oder diskret (vgl. [23])?
- *Gibt es eine Kontrollkondition, um einen Vergleich zwischen verschiedenen Visualisierungsmethoden zu erleichtern* (vgl. [15])*?*
- Wie werden die Messergebnisse analysiert? Der Datentyp von abhängigen Variablen hat einen Einfluss auf die Wahl der statistischen Auswertungsmethoden (vgl. [3, 23]). Außerdem ist die Anzahl von unabhängigen und abhängigen Variablen in dieser Phase bekannt, dadurch lassen sich die passenden Methoden von unpassenden früh abgrenzen.
- Wie viele Versuchspersonen sind zu erwarten/müssen teilnehmen?
- Welche Qualifikation ist bei den Versuchspersonen notwendig/zu erwarten?
- Prüfung: Können Auswahleffekte vermieden werden?
- Prüfung: Ist eine Beeinflussung durch die Anforderungscharakteristik zu befürchten?
- Gibt es Methoden zur Identifizierung von unbrauchbaren Antworten (vgl. [15])? Können/müssen diese im Rahmen dieser Studie entwickelt werden?
- Prüfung: Falls die Antworten für alle Aspekte dieser Phase gefunden wurden, lässt sich damit tatsächlich eine relevante Forschungsfrage beantworten?

In der Phase der **Implementierung** wird die konkrete Gestaltung des Experiments definiert. Viele wichtige Facetten des Experimententwurfs und oft sogar der behandelten Fragestellung lassen sich erst in dieser Phase klären.

- Muss das Wissen/Vorwissen der Teilnehmer geprüft werden?
- Ist ein Vortest nötig/möglich?
- Ist ein vorheriges Training nötig/möglich (vgl. [15, 20])?
- Müssen die Teilnehmer auf Stereo-Fähigkeit geprüft werden?
- *Müssen die Teilnehmer auf adäquate räumliche Sehschärfe und Farbenblindheit geprüft werden* (vgl. [15])*? Diese und die vorherige Frage sind wichtig, um unbrauchbare Antworten zu verhindern. Es ist nicht ausgeschlossen, dass ein Proband sehr schlecht sieht oder kaum hört und nur des Geldes wegen an der Studie teilnimmt.*
- Wie werden die Aufgaben den Teilnehmern präsentiert?

  Es ist sehr wichtig, die Versuchsunterlagen in schriftlicher Form zu haben: Zum einen kann der Versuchsleiter seine mündliche Erklärung nicht immer identisch bei jedem Versuchsdurchlauf wiedergeben, zum anderen ist es möglich, dass die Probanden sich so nur einen Teil der mündlichen Hinweise merken oder unterschiedliche Schwerpunkte bei Hinweisen setzen, um die Aufgabe zu erfüllen (vgl. [15]).

- *Können Lern- und Reihenfolgeneffekte auftreten* (vgl. [15, 20])*?*
- *Müssen die Stimuli in zufälliger Reihenfolge präsentiert werden, um Lern- und Reihenfolgeneffekte zu verhindern* (vgl. [15])*?*
- *Wie können Lern- und Reihenfolgeneffekte sonst noch verhindert werden?*
- Sollen bei der Präsentation der Stimuli eine oder mehrere Pausen gemacht werden (vgl. [15])? An welchem Zeitpunkt sollen sie eingeplant werden und wie lange sollen sie dauern?
- Wie kann die Sterblichkeit vermieden werden?
- Welche Infrastruktur benutzen die Teilnehmer bei ihrer Arbeit? (Versuchsumgebung)
- Prüfung: Lässt sich durch die Beachtung der genannten Aspekte eine ausreichende innere Gültigkeit sicherstellen?
- Prüfung: Ist auch eine ausreichende äußere Gültigkeit zu erwarten?

In der Phase des **Tests** muss der Experimentaufbau überprüft und seine Schwächen korrigiert werden. Dafür sollte eine Pilotstudie mit einigen Teilnehmern durchgeführt werden. Die Durchführung einer Pilotstudie kann helfen, Kosten und Zeit zu sparen. Das Ziel von Pilotstudien ist es, Fehler und Probleme im Design aufzudecken (vgl. [15]). Folgende Fragen sind zu klären:

- Sind die Versuchsunterlagen vollständig?
- Sind die Anweisungen in den Versuchsunterlagen verständlich und eindeutig (vgl. [15, 20])?
- Funktioniert die Versuchsumgebung (aus Teilnehmersicht)?
- Ist die Versuchsumgebung angemessen?
- Funktioniert die Messapparatur?
- Ist der Schwierigkeitsgrad der Aufgaben angemessen (vgl. [15, 20])?
- Ist die ausreichende äußere Gültigkeit zu erwarten?

Die **Ausführungsphase** besteht aus drei Komponenten: dem Anwerben und Vorbereiten von Versuchspersonen, der Ausführung im engeren Sinne und den begleitenden Maßnahmen, die ein Fehlschlagen des Experiments verhindern sollen. Dabei sind folgende Gesichtspunkte wichtig:

- Wie motiviere ich Versuchspersonen zur Teilnahme? Prüfung: Entstehen dabei Auswahleffekte?

  Meistens werden die Probanden durch die finanzielle Belohnung zur Teilnahme motiviert, oft kann aber die Neugier auf neue wissenschaftliche Methoden und Technologien deren Beweggrund sein. So ist es wichtig, eine interessante Anzeige mit der Beschreibung der Methode und der Fragestellung, die untersucht werden soll, zu erstellen.

- Sind alle Teilnehmer ausreichend bereit und fähig, die Aufgabe zu erfüllen (vgl. [15, 20])?
- Wie behalte ich die Übersicht über den Fortgang des Experiments?
- Prüfung: Funktioniert die Messapparatur wirklich in jedem Einzelfall?
- Prüfung: Bedroht unvorhergesehenes Verhalten der Teilnehmer das Experiment?
- Wie vermeide ich Sterblichkeit? Prüfung: Entsteht dabei eine Verzerrung? Wie verhindere ich Lawineneffekte beim Ausscheiden von Teilnehmern?
- Wie vermeide ich, dass künftige Teilnehmer von früheren Telnehmern beeinflusst werden (insbesondere die Lösungen erfahren)?

In der Phase der **Auswertung** soll zuerst die Validierung der Daten erfolgen, danach findet ihre Beurteilung im Hinblick auf die Experimentfrage statt und zum Schluss die Analyse für Zusatznutzen oder Milderung von Problemen.

- Prüfung: Sind die gesammelten Daten vollständig?
- Prüfung: Sind in den Daten Inkonsistenzen zu entdecken?
- Prüfung: Sind die Daten unglaubwürdig?
- Welche Ergebnisse im Hinblick auf die Experimentfrage sind den Daten zu entnehmen?
- *Was ist die Ursache dafür, dass die Ergebnisse bei einer Visualisierung in der Zeit besser aber in der Genauigkeit schlechter als bei einer anderen sind? War die Visualisierung wirklich schneller oder wurden die Teilnehmer müde und haben nur geraten* (vgl. [18]) *?*
- Welche Bedrohungen der inneren Gültigkeit sind zu erkennen? Können diese ausgeräumt werden?
- Können die Beobachtungen erklärt werden?
- Welche sonstigen unerwarteten Beobachtungen gibt es?

Ein großes Problem bei einer Studie ist ein zweifelhaftes Ergebnis. In der Regel bedeutet das, dass es Design-Fehler in der Studie gab und sie erneut durchgeführt werden muss. Normalerweise ist dabei nur ein Teil der Studie betroffen, so dass der Aufwand bei der zweiten Durchführung wesentlich geringer ist. Bei einer Wiederholung ist es möglich, zusätzliche Hypothesen zu testen, falls diese aus erfolgreichen Teilen der Studie entstanden sind (vgl. [15]).

Die Experimentergebnisse sollten in jedem Fall schriftlich dokumentiert werden. Sie können in der Phase der **Publikation** entweder öffentlich sichtbar gemacht werden, indem sie in einer wissenschaftlichen Zeitschrift erscheinen oder nur einem kleinen Kreis von Lesern zugänglich sein – die Anforderungen bleiben in beiden Fällen ähnlich:

- Welches Publikum sollte ich über meine Ergebnisse informieren?

- Wie überzeuge ich es davon, dass das Experiment relevant ist?
- Wie überzeuge ich es davon, dass das Experiment glaubwürdig ist?
- Wie stelle ich den Experimentaufbau knapp und klar dar?
- Auf welche Bedrohungen der inneren Gültigkeit muss ich hinweisen?
- Wie stelle ich die Ergebnisse genau und unmissverständlich dar?
- Wo und in welcher Form stelle ich für spätere Metastudien die rohen Ergebnisdaten bereit?
- Auf welche Beschränkungen der äußeren Gültigkeit muss ich hinweisen? Welche Spekulationen über äußere Gültigkeit sind angemessen?

Es ist eine große Herausforderung, ein gutes Design für eine Vergleichsstudie im Visualisierungsbereich zu entwerfen. Dabei müssen sehr viele Aspekte beachtet werden. Die oben erstellte Liste mit grundlegenden Fragen sollte bei dem Entwurf und der Durchführung einer Studie hilfreich sein.

3.2 Implementierung, Durchführung und Auswertung

In diesem Abschnitt werden weitere Aspekte beschrieben, die während der Implementierung sowie während der Durchführung und der Auswertung beachtet werden müssen. Die erwähnten Aspekte sind in Anlehnung an Prechelt [20] dargestellt.

3.2.1 Implementierung und Durchführung

Nachdem der Experimententwurf vollendet ist und die Versuchspersonen bekannt sind, sollen der Experimentaufbau realisiert und das Experiment durchgeführt werden. Das Wichtigste in beiden Phasen ist es zu verhindern, dass das Experiment fehlschlägt. Ein solches Fehlschlagen könnte schon durch einen ungeeigneten Experimententwurf vorprogrammiert sein. Ist dies nicht der Fall, müssen die folgenden beiden Störungen vermieden werden:

- Die Schwierigkeiten, die nicht mit Experimentfragen oder eigentlichen Aufgaben zu tun haben (Verwirrung), könnten zur Verschlechterung der Leistung oder dem Abbruch des Experiments führen.
- Durch Messversagen werden die abhängigen Variablen nicht zuverlässig oder genau genug erfasst, so dass nicht alle Daten für die Auswertung vorhanden sind.

Da moderne Experimente häufig sowohl mit Menschen als auch mit Computern arbeiten, kann bei ihrer Durchführung viel Unvorhersehbares passieren. Folglich soll beim Experimentaufbau

die maximale Robustheit erstrebt werden. Sie wird durch Orientierung an den zwei Prinzipien Einfachheit und Redundanz erreicht.

Es existiert inhaltliche, strukturelle oder technologische **Einfachheit**. Inhaltliche und strukturelle Einfachheit wird durch eine eingängige und klare Formulierung von Aufgabenstellungen erreicht. So ist gesichert, dass die Versuchspersonen nichts übersehen oder falsch zuordnen können. Technologische Einfachheit bedeutet, komplizierte (und daher anfällige) Messaufbauten wie auch komplexe Arbeitsumgebungen und Hilfsmittel zu meiden, damit die Versuchspersonen nicht überfordert werden. **Redundanz** in der Aufgabenstellung kann helfen, Verwirrung oder Versehen zu verhindern. Da es sowohl bei einfacher als auch bei komplizierter Messanordnung unmöglich ist, das Messversagen ganz auszuschließen, ist es optimal, wenn für die Erfassung der wichtigsten Daten alternative Verfahren verwendet werden.

3.2.2 Auswertung der Beobachtungen

In dieser Phase gelten zwei Grundprinzipien, die ähnlich der Einfachheit und der Redundanz der Implementierung sind:

- Bevorzuge einfache Auswertungsmethoden gegenüber komplizierten.
- Bevorzuge anschauliche Methoden gegenüber abstrakten.

Das dritte Prinzip in der Phase heißt:

- Nutze stets graphische Methoden, um die Analysen oder ihre Ergebnisse zu veranschaulichen.

Die Anwendung von statistischen Methoden ist nicht einfach und wird durch die Tatsache erschwert, dass die quantitativen Daten nicht nur interessante, sondern auch überraschende und tückische Eigenschaften besitzen. Besteht bei der Auswertung der Daten die Möglichkeit, die Hilfe eines professionellen Statistikers in Anspruch zu nehmen, sollte sie unbedingt genutzt werden. Meistens ist diese Möglichkeit aber nicht vorhanden, daher müssen die einfachen Auswertungsmethoden bevorzugt werden. Dadurch bleibt die Anzahl der Fehlerquellen klein. Zusätzlich soll die Bedeutung der Methoden leicht nachvollziehbar sein, damit sich eventuelle Fehler und Schwächen der Analyse leichter entdecken lassen. Wichtig ist in dieser Phase zu beachten, dass die Grenze zwischen Schwäche und Fehler in der Statistik fließend, sogar vage ist, „weil es völlig korrekte Anwendungsfälle nur selten gibt“ [20].

Bevor die gesammelten Daten einem Statistikprogramm übergeben werden, sollen sie auf Konsistenz und Glaubwürdigkeit geprüft werden. Folgende Fragen sind für eine **Konsistenzprüfung** zu klären:

- Fehlen Datenwerte bei Variablen, für die das nicht sein kann?
- Sind Werte negativ, die das nicht sein können?

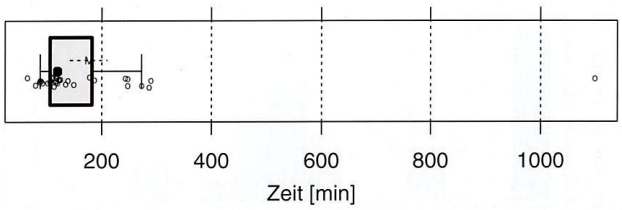

Abbildung 8: Entdeckung von Eingabefehlern durch eindimensionalen Punktplot (kleine Kringel): Der Wert 1099 sollte eigentlich 109 sein [20]

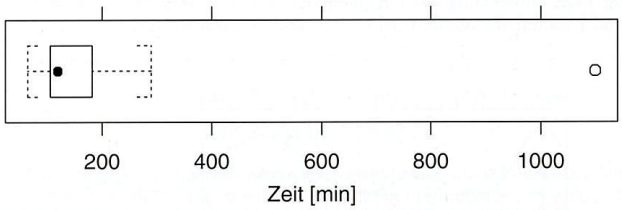

Abbildung 9: Entdeckung von Eingabefehlern durch herkömmlichen Boxplot: (Werte jenseits von 1,5 Boxbreiten außerhalb der Box werden als mögliche Ausreißer separat angezeigt): Der Wert 1099 sollte eigentlich 109 sein [20]

- Sind Werte Null, die das nicht sein können?
- Sind Werte größer als möglich? Beispiel: Prozentwerte größer als 100 oder Zeitdauern größer als die Gesamtdauer des Experiments.
- Gibt es bei Variablen mit Aufzählungstyp unerwartete Werte? Beispiel: ein falsch geschriebener Name oder ein falscher Gruppenname.
- Sind alle Konsistenzbedingungen zwischen mehreren Variablen erfüllt? Beispiel: Ist die Anzahl gegebener Antworten kleiner als die Anzahl korrekter Antworten?

Diese Fragen dienen dem Entdecken von Unregelmäßigkeiten in den Daten. Hierbei gilt: je mehr Redundanz in den gesammelten Daten vorhanden ist, desto besser lassen sich die möglichen Fehler auffinden.

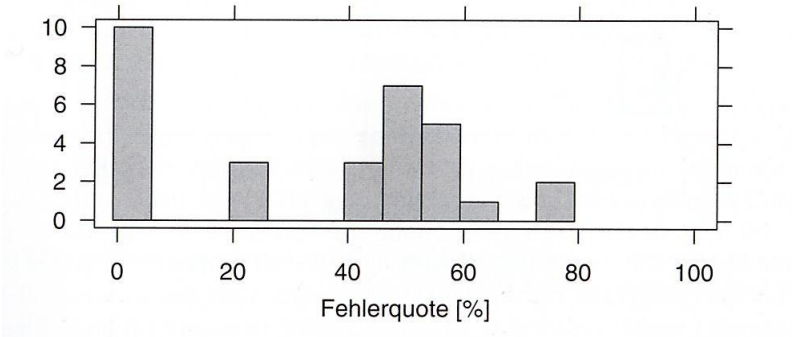

Abbildung 10: Entdeckung von Eingabefehlern durch ein Histogramm: Die zahlreichen Werte zwischen 0 und 1 sollten alle hundertmal so groß sein (Prozentwerte) [20]

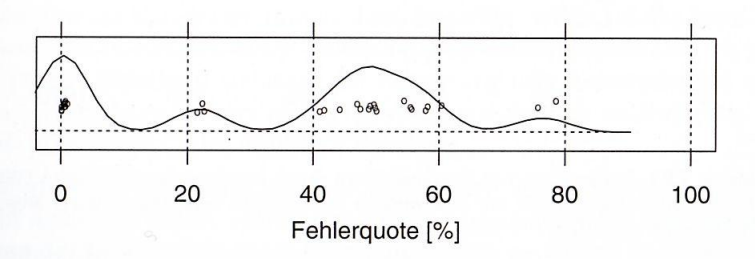

Abbildung 11: Entdeckung von Eingabefehlern durch Dichteplots; die Kurve gibt die Wahrscheinlichkeitsdichte an. Die zahlreichen Werte zwischen 0 und 1 sollten alle hundertmal so groß sein (Prozentwerte). Die Daten entsprechen denen von Abbildung 10 [20]

Der nächste Schritt nach der Konsistenzprüfung ist die **Glaubwürdigkeitsprüfung**. „Glaubwürdigkeit prüfen heißt, wahrscheinliche Eigenschaften der Daten zu testen, um unplausible Daten zu entdecken. Diese sind manchmal korrekt, oft aber falsch“ [20].

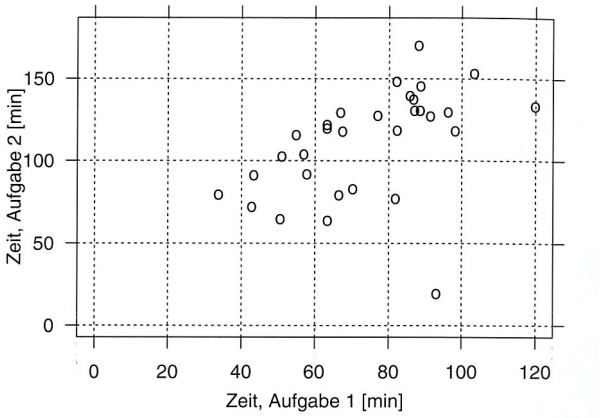

Abbildung 12: Entdeckung von Eingabefehlern durch zweidimensionale Punktplots. Wer bei Teilaufgabe 1 (x-Achse) recht langsam war, ist vermutlich nicht bei Teilaufgabe 2 (y-Achse) besonders schnell: Die Werte beim Punkt (93,19) sollten eigentlich bei (93,91) liegen [20]

Abbildung 13: Entdeckung von Eingabefehlern durch eindimensionale Plots von Quotienten. Der Datenpunkt bei 4,9 sollte eigentlich bei 1,0 liegen [20]

Für eine einzelne Variable ist zu beachten:

- Kommen wenige ungewöhnlich hohe Werte vor?

  Mögliche Fehler: Eingabefehler (z. B. 1099 anstatt 109) oder Versagen einer automatischen Messung. Geeignete Suchhilfsmittel: Eindimensionale Punktplots (Abb. 8) oder herkömmliche Boxplots, in denen „Ausreißer“ mit Punkten dargestellt werden (Abb. 9).

„Boxplots ... dienen zur groben Darstellung der Verteilung von Werten einer Stichprobe. Der Kasten gibt den Bereich an, in dem die mittlere Hälfte der Daten liegt, d. h. ein Viertel liegt links vom Kasten, zwei Viertel im Kasten und ein Viertel rechts davon. Der Kastenrand wird nötigenfalls zwischen zwei Datenpunkten interpoliert. Die Breite des Kastens heißt *Interquartilbereich*. Die Markierung im Kasten kennzeichnet den *Median*, also die Grenze zwischen der oberen und unteren Hälfte der Daten. Die Schwänze rechts und links des Kastens geben bei herkömmlichen Boxplots üblicherweise den äußersten Datenpunkt an, der noch innerhalb von 1,5 Interquartilbereichen außerhalb des Kastens liegt. Alle noch weiter entfernt liegenden Datenpunkte werden als *Ausreißer* betrachtet und entweder durch separate einzelne Punkte im Plot angegeben oder ganz unterdrückt. Boxplots ... eignen sich insbesondere zum schnellen visuellen Vergleich mehrerer etwa gleich großer Stichproben. Die beschriebene Handhabung der Schwänze und Ausreißer ist nur für Stichproben mit wenigen Datenpunkten sinnvoll“ [20].

- Kommen wenige ungewöhnlich niedrige Werte vor?

  Mögliche Fehler: Eingabefehler (z. B. 19 anstatt 109) oder Versagen einer automatischen Messung. Geeignete Suchhilfsmittel: wie oben.

- Kommen viele ungewöhnlich hohe oder niedrige Werte vor?

  Mögliche Fehler: Wechsel der Maßeinheit während der Dateneingabe, z.B. Stunden versus Minuten oder 0 bis 1 versus Prozent. Geeignete Suchhilfsmittel: Histogramme (Abb. 10) oder Dichteplots (Abb. 11)

- Kommt ein Wert ungewöhnlich häufig vor?

  Mögliche Fehler: Versagen einer automatischen Messvorrichtung. Geeignete Suchhilfsmittel: Histogramme, Dichteplots, eindimensionale Punktplots.

Betrachte Zusammenhänge zwischen zwei Variablen:

- Kommen unwahrscheinliche Kombinationen vor?

  Mögliche Fehler: Eingabefehler (z. B. 19 anstatt 109) oder Versagen einer automatischen Messung. Geeignete Suchhilfsmittel: zweidimensionale Punktplots (Abb. 12), eindimensionale Plots von Quotienten (Abb. 13).

Falls eine Glaubwürdigkeitsprüfung mehr als eine kleine Zahl von Fehlern aufdeckt, sind viele weitere vermutlich unentdeckt geblieben. Dann sollte die Dateneingabe bzw. die Messung noch einmal sorgfältiger wiederholt werden.

4 Eye-Tracking-Studie

4.1 Phase der Anforderungsbestimmung

Zu Beginn der Durchführung einer Studie findet die Phase der Anforderungsbestimmung statt. Hier soll die Fragestellung gewählt werden, die in der Studie untersucht wird. Da das Thema und die Randbedingungen der Diplomarbeit von Anfang an festgelegt waren, wurden dadurch die meisten Fragen zu dieser Phase beantwortet.

Das Ziel der Studie bestand darin, einige Hierarchievisualisierungen zu vergleichen. Zeitgrenzen, finanzielle Mittel, Geräte, Räume und die Anzahl der Versuchsleiter waren von vornherein bekannt. Die Prüfung dieser Randbedingungen hat gezeigt, dass sie für die Lösung der Aufgabenstellung akzeptabel sind.

Für den Vergleich verschiedener Subgruppen wurden keine statistischen Tests eingeplant. Die Suche nach eventuellen Unterschieden zwischen Geschlechtern oder Altersgruppen sollte anhand von Heat Maps und Gaze Plots durchgeführt werden. Die Durchführung eines solchen Vergleichs hing davon ab, ob und wie gut die Altersgruppen bzw. Geschlechter in der Menge der Probanden präsentiert sind.

In der Auswertungsphase sollten die gesammelten Daten anhand von Heat Maps und Gaze Plots auf die Strategien untersucht werden, die die Probanden möglicherweise für die Lösung der Aufgaben entwickeln.

4.2 Entwurfsphase

In der Phase des Entwurfs soll die grobe Struktur des Experiments festgelegt werden. Die wichtigsten Fragen in dieser Phase sind mit den Messaspekten, dem Typ der Daten und der Anzahl der Variablen verbunden.

Am Anfang der Entwurfsphase wurde festgelegt, welche konkreten Hierarchievisualisierungen verglichen werden. Es gab die Überlegung Tree-Map- und Icicledarstellungen in die Studie mit zu integrieren. Letztendlich fiel die Entscheidung, nur Baumdiagramme zu untersuchen und nur eine Testaufgabe zu stellen: Der Proband soll den kleinsten gemeinsamen Vorfahrknoten aller rot markierten Blätter finden.

Danach wurden die zu variierenden Aspekte der Darstellungsart der Bäume bestimmt. Der erste Aspekt war der Typ der Bäume (traditionell, orthogonal und radial), der zweite Aspekt war die Lage der Wurzel (oben, unten, links und rechts). Die Anzahl der markierten Blätter (3, 6 und 9), die Anzahl der Levels (10) und die Anzahl der Knoten (500-600) sollten nicht als unabhängige Variable betrachtet werden. Zur abhängigen Variablen wurde die Zeit, die vergeht, bis der Proband die Lösung findet.

Geplant waren die zwei Vergleiche der Ergebnisdaten. Es sollten radiale, traditionelle und orthogonale Bäume verglichen werden, wobei von den letzten beiden die Darstellung mit der Wurzel oben genommen werden sollte. Im ersten Vergleich hat sich der Typ der Bäume als die einzige

unabhängige Variable ergeben. Der zweite Vergleich sollte inmitten von traditionellen bzw. orthogonalen Darstellungstypen durchgeführt werden: Die Darstellung mit der Wurzel oben (bzw. rechts) sollte mit der Darstellung mit der Wurzel unten (bzw. links) verglichen werden. Diese Aufteilung war deswegen notwendig, weil die Darstellungen mit Wurzel links/rechts auf die Bildschirmgröße skaliert werden mussten und so bei der Präsentation im eigentlichen Test kleiner als die Dartellungen mit Wurzel oben/unten waren. Beim zweiten Vergleich wurde lediglich die Lage der Wurzel zur unabhängigen Variablen. Die zu messende abhängige Variable wurde in den beiden Fällen die Zeit, die verstreicht, bis der Proband die Lösung gefunden hat.

Die Messung der Zeit sollte mit dem Eye-Tracker-Gerät durchgeführt werden. Mehrere Probedurchläufe haben gezeigt, dass das Gerät gut und zuverlässig funktioniert. Der einzige mögliche Fehler bestand darin, dass der Eye-Tracker wegen unsauberer Brillen oder langem Blinzeln die Pupillen kurzfristig nicht verfolgen konnte. Dadurch splittete sich die Aufnahme für einen Proband und einen Testdurchlauf in zwei auf. Dieser Fehler trat aber sehr selten auf, daher wurden die möglichen Messfehler in einem 3-25-prozentigem Bereich erwartet.

Das Messen mittels des Eye-Trackers lieferte quantitative Daten und hat gewährleistet, dass das, was gemessen werden soll (die Zeit) auch gemessen wird.

Da der Eye-Tracker bei mehr als einer Aufnahme für einen bestimmten Stimulus die Zeit mit der Genauigkeit von zwei Nachkommastellen anzeigt, wurde eine diskrete abhängige Variable gemessen. Zu diesem Typ der Variable passt eine Verhältnisskala. Als geeignete Auswertungsverfahren für die zu messenden Daten haben sich die einfaktorielle ANOVA und der t-Test erwiesen.

Sehr früh war klar, dass in der Studie der Einsatz von Testdaten notwendig ist und dass es synthetische Daten sein werden. Die zu präsentierenden Bäume sollten vor der Studie zufällig generiert und auf Fehler und Inkonsitenzen überprüft werden. Ihre endgültige Anzahl wurde erst nach dem Beenden der Pilotstudie festgelegt.

Es wurde keine Kontrollkondition gefunden, die den Vergleich zwischen verschiedenen Visualisierungsarten erleichtern könnte.

In der Entwurfsphase wurde geplant, für die Teilnahme an der Studie zwischen 30 und 45 Teilnehmer zu finden. Bei dieser Anzahl hält sich einerseits der Aufwand in Grenzen, andererseits sind die Replikation, Randomisierung und die Signifikanz bei der Auswertung der Ergebnisse gewährleistet.

Da keine Vorkenntnisse bzw. Qualifikation für die Teilnahme an der Studie notwendig waren, bedrohten auch keine möglichen Auswahleffekte die innere Gültigkeit der Studie. Ebenso gab es keinen Anlass, die Beeinflussung der Anforderungscharakteristik zu befürchten, weil von keiner Gruppe der Probanden besonders gute bzw. besonders schlechte Leistungen erwartet wurden.

Um unbrauchbare Antworten zu identifizieren, wurden Seh- und Farbtests eingeplant. Bei der Durchführung dieser Tests ist es wichtig, die richtige Größe von Buchstaben bzw. von Bildern und den richtigen Abstand zur Wand bzw. zum Bildschirm einzustellen und einzuhalten. Die Buchstaben und Bilder müssen auf Augenhöhe der Teilnehmer präsentiert werden.

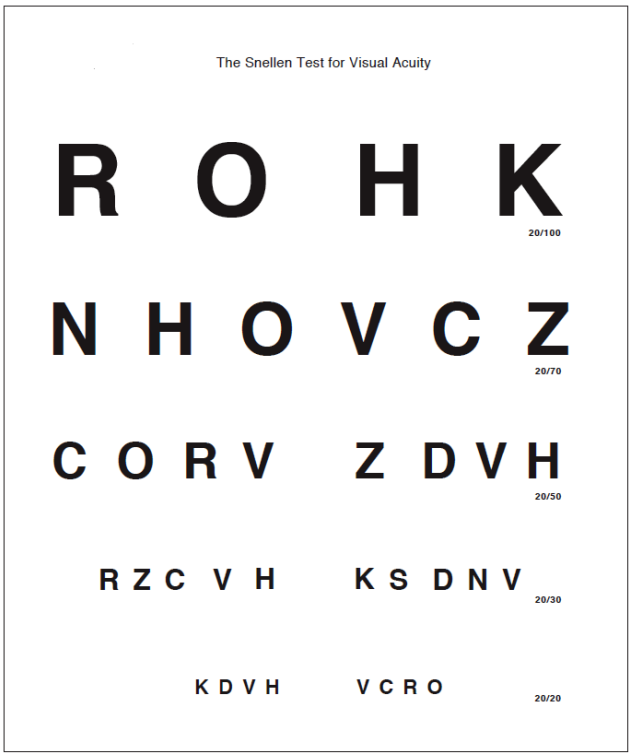

Abbildung 14: Der Sehtest von Snellen

Für das Testen des Sehvermögens wurde der Sehtest vom niederländischen Augenarzt Herman Snellen (1834 – 1908) verwendet (Abb. 14). Um die Teilnehmer auf die Farbenblindheit zu prüfen, wurde der Test vom japanischen Arzt Ishihara Shinobu (1879 - 1963) eingesetzt (Abb. 15).

In einer Studie sind diese Tests dafür da, um minimal notwendige Fähigkeiten der Versuchspersonen festzustellen. Sie sind einer medizinischen Untersuchung nicht äquivalent. Sollten mögliche Hinweise auf Farbenblindheit vorhanden sein, darf der Proband darüber nicht informiert werden und soll die Studie zu Ende führen. Je nach Rolle der Farbe im Experiment, den verwendeten Farben und dem Fehlerniveau in den Ergebnissen (die Lage über oder unter dem Fehlerdurchschnitt), sollte entschieden werden, ob der Datensatz verworfen wird oder nicht.

In der Entwurfsphase wurde keine Entwicklung der anderen speziellen Maßnahmen zur Identifizierung der unbrauchbaren Antworten eingeplant und entsprechend der Planung auch keine durchgeführt.

4.3 Phase der Implementierung

Nachdem die Fragen des Entwurfs beantwortet sind, müssen in der Phase der Implementierung konkrete Facetten des Studienablaufs geklärt werden. Es geht in dieser Phase um die Details, die mit der möglichen Prüfung der Probanden und mit den Einzelheiten der Aufgabenstellung verbunden sind.

In der durchgeführten Studie sollte das Wissen und Vorwissen der Probanden nicht geprüft

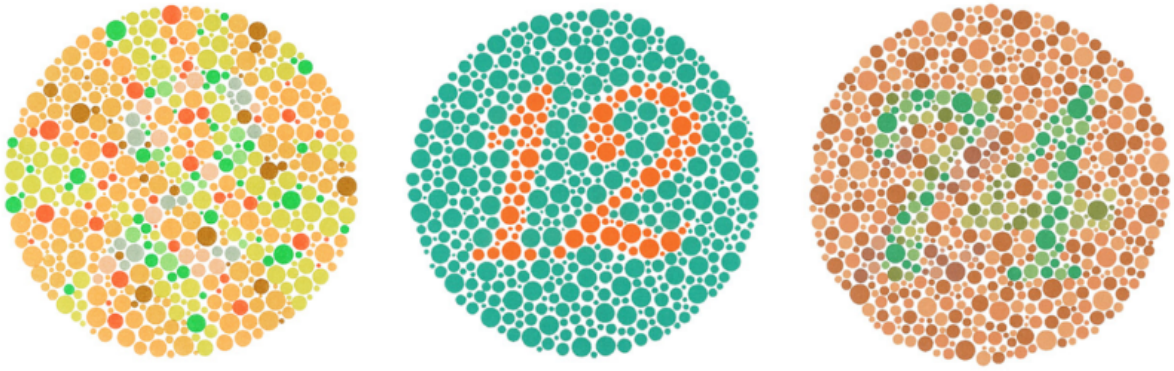

Abbildung 15: Einige der von Ishihara verwendeten Farbtafeln

werden, es wurde auch kein Vortest eingeplant. Für die Erklärung der Aufgabenstellung wurde sowohl eine kompaktere als auch eine ausführlichere Form entwickelt, abhängig davon, ob die Probanden schon mit den Baumdiagrammen zu tun hatten oder nicht.

Ein vorheriges Training auf den gedruckten Beispielbäumen und ein kleiner Probedurchlauf am Eye-Tracker wurden in den Testablauf eingeplant. So sollte sichergestellt werden, dass die Teilnehmer sowohl die notwendigen Begriffe als auch die Aufgabenstellung richtig verstanden haben. Der Probedurchlauf am Eye-Tracker war wichtig, um einerseits die mögliche Nervosität der Teilnehmer abzubauen und andererseits die Schwierigkeiten, die nicht mit der eigentlichen Aufgabe zu tun haben, zu verhindern.

Es wurde kein Test der Stereofähigkeit eingeplant. Solch ein Test ist in der Begrüßung und den ersten organisatorischen Hinweisen implizit enthalten und reicht für die konkrete Erfüllung der Aufgabenstellung aus.

Die Erklärung für die im Test verwendeten Begriffe befand sich in einem schriftlichen Tutorial. Die eigentliche Aufgabenstellung, nach dem kleinsten gemeinsamen Vorfahrknoten aller rot markierten Blätter zu suchen, wurde ebenso in das Tutorial eingefügt.

Die Baumdarstellungen wurden in die drei Blöcke mit jeweils traditionellen, orthogonalen und radialen Bäumen aufgeteilt. Es war klar, dass bei der konstanten Reihenfolge von Stimuli Lern- und Reihenfolgeeffekte auftreten werden. Um diese zu vermeiden, wurde eingeplant, die drei Blöcke durch die manuelle Wahl des Testblocks zu permutieren. Inmitten von den Blöcken wurde zusätzlich die Permutationsfunktion angewendet, die der Eye-Tracker dafür anbietet.

Nach jedem Testblock sollte dem Proband die Möglichkeit gegeben werden, eine maximal fünf Minuten lange Pause zu machen. Eine feste Pause bestimmter Zeit wurde für jeden Testdurchlauf nicht eingeplant.

Die Sterblichkeit in der Studie war sehr wahrscheinlich bei Probanden, die nicht bezahlt werden konnten. Zu dieser Gruppe gehörten alle Mitarbeiter und Hilfswissenschaftler der Universität Stuttgart. Um die Sterblichkeit aus diesem Grund zu verhindern, wurde in der allgemeinen Erinnerungsmail ein Hinweis über die Bezahlungsbedingungen eingefügt. Diese Mail sollte immer zwei Tage vor dem jeweiligen Testtermin geschickt werden. Nur eine einzige Kandidatin hat wegen dieser Bedingung abgesagt.

Einer der ersten Probanden hatte leichte Schwierigkeiten während der Lösung der Aufgabe bei den Baumdarstellungen mit der Wurzel links bzw. rechts. Das geschah deswegen, weil er nur bei kleineren Schriften für die Arbeit am Bildschirm seine Brille benutzte und diese zum Test nicht mitgenommen hat. Aus diesem Grund enthielt die Erinnerungsmail eine zusätzliche Bemerkung für alle Brillenträger, dass sie ihre Brillen nicht vergessen.

Der Eye-Tacker befand sich in einem ca. 30 Quadratmeter großen Labor. Die Infrastruktur, die die Teilnehmer bei ihrer Arbeit benutzen mussten, begrenzte sich auf den Eye-Tracker-Bildschirm und die Maus.

Die Prüfung der möglichen Bedrohungen für innere Gültigkeit des Experiments hat gezeigt, dass alle zu diesem Zeitpunkt bekannten Problemaspekte beachtet wurden und dass die entsprechenden Gegenmaßnahmen getroffen wurden.

4.4 Phase des Tests (Pilotstudie)

In der Phase des Tests soll eine Pilotstudie durchgeführt werden. Dadurch lassen sich die meisten Aspekte des Entwurfs und der Implementierung überprüfen. Sehr viele Fehler und Schwächen können auf diese Weise aufgedeckt und ausgeräumt werden. Oft laufen in der Praxis einige Abschnitte des Tests anders ab, als es vorher angenommen wird. Deswegen ist es sehr wichtig, einige Probedurchläufe durchzuführen, um die wertvolle Erfahrung zu sammeln.

In der Pilotstudie kann der Versuchsleiter vieles dazulernen und eigene Nervosität abbauen. So tritt er später viel sicherer in der eigentlichen Studie auf und hat ihren Ablauf unter Kontrolle. Die Pilotstudie wurde mit fünf Teilnehmern durchgeführt, wobei die Zahl als optimal angenommen werden kann. Bei den ersten zwei Teilnehmern sind noch sehr viele Fehler gemacht worden. Die letzten drei Durchläufe wurden benutzt, um die restlichen Schwächen zu beheben und den Studienverlauf zu optimieren. Beim fünften Teilnehmer gab es schon praktisch keinen Anlass mehr für Veränderungen oder Verbesserungen.

Für die Anzahl der Teilnehmer ist die fünf optimal deswegen, weil bei weniger als 4 Teilnehmern die Gefahr besteht, dass noch viele Fehler unentdeckt bleiben. Ab 7 Teilnehmer kann der Aufwand unnötig ansteigen, ohne irgendeinen Ertrag zu bringen. Wobei die Zahl der Probanden in einer Pilotstudie selbstverständlich von jedem konkreten Fall abhängt. Am Ende der Pilotstudie könnten die Experimentatoren flexibel sein und je nach dem Verlauf zusätzliche Teilnehmer einladen bzw. den “überflüssigen” Probanden absagen. Da die Teilnehmer in einer Pilotstudie normalerweise nicht bezahlt werden und eher als Helfer zur Verfügung stehen, soll solch eine Absage kein Problem für die beiden Seiten darstellen.

In der durchgeführten Pilotstudie gab es keine Probleme mit der Messapparatur. Die Versuchsunterlagen waren vollständig, verständlich und eindeutig. Die Unterlagen wurden nur an einigen Stellen minimal verbessert. Der Schwierigkeitsgrad der Aufgaben und die Anzahl der Stimuli haben sich als optimal gezeigt. Die Testumgebung funktionierte aus Teilnehmersicht gut und hat sich als angemessen erwiesen.

Wichtig war es in der Pilotstudie, das konstante Experimentatorverhalten und den konstanten Erklärungsablauf zu trainieren. So konnte verhindert werden, dass diese Aspekte zu einer Störvariable beim Durchlauf der eigentlichen Studie werden. Schon vor der Pilotstudie wurden die Sätze entworfen, die gesagt werden sollen. Während der ganzen Pilotstudie wurden sie verbessert, vervollständig und ihre Reihenfolge wurde immer wieder optimiert.

Abbildung 16: Plakat zum Aufruf der Studie

4.5 Ausführungsphase

Da die Pilotstudie gut verlief und erfolgreich zu Ende ging, wurde gleich nach ihrem Abschluss die Ausführungsphase begonnen. Um die Versuchspersonen anzuwerben, wurde eine Rundmail an 30 Teilnehmer einer früheren Studie geschickt und ein Plakat an 5 Studentenwohnheimen sowie in dem Informatikgebäude ausgehängt (Abb. 16).

Nach 4 Tagen haben sich aber nur 4 Teilnehmer gemeldet. Da die Studie schon in einer Woche starten sollte, wurden zügig noch viele weitere Aushänge an der Hochschule für Druck und Medien und in der Sportanlage der Universität Stuttgart angebracht. Eine Rundmail wurde an alle Studenten der Fakultät Informatik und noch an ca. 7 Teilnehmer einer parallel laufenden Studie geschickt. Insgesamt haben sich über 45 Interessierte gemeldet und 38 haben an der Studie teilgenommen.

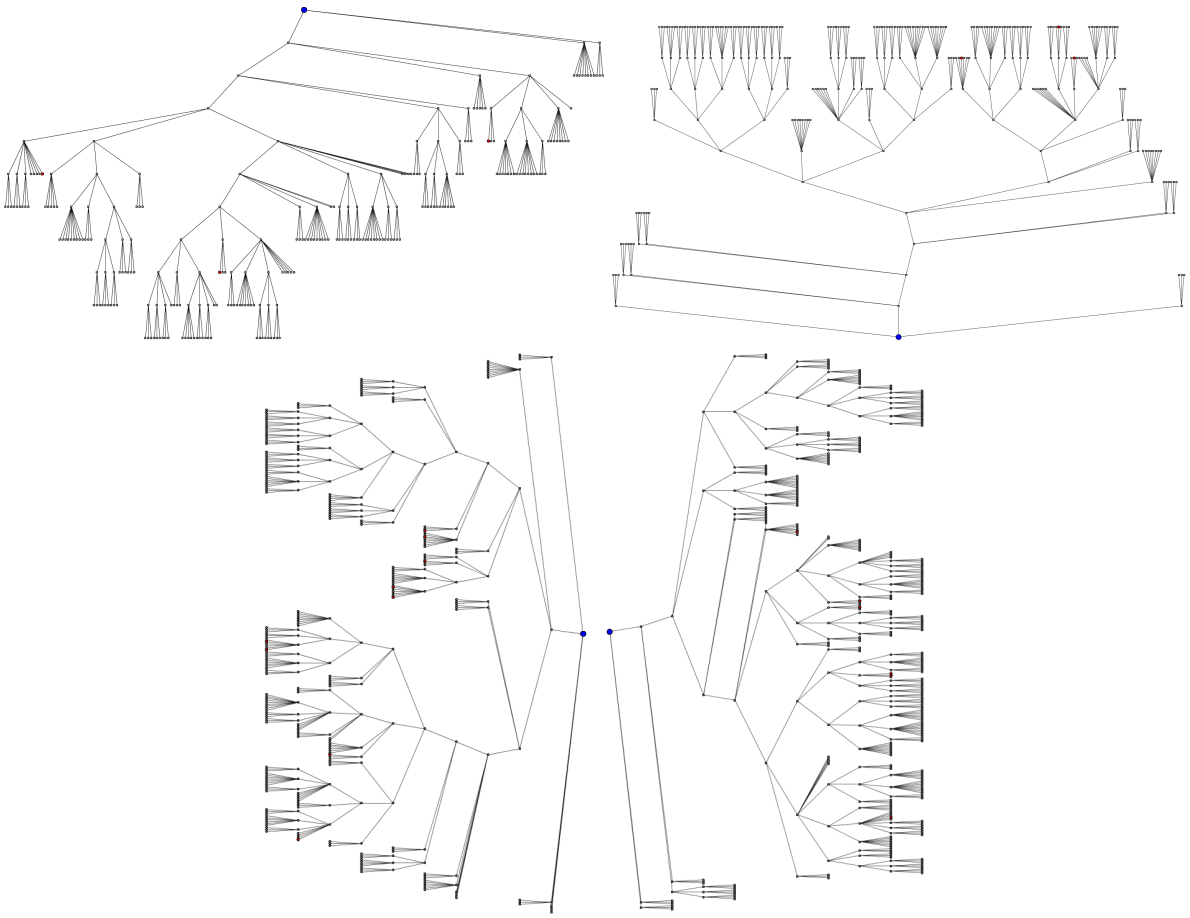

Abbildung 17: Beispiele der vier Varianten für die traditionelle Darstellungsart

4.5.1 Probanden

In der Studie nahmen 10 weibliche und 28 männliche Probanden teil. Deren Durchschnittsalter betrug 24,6 Jahre. Die zwei jüngsten Probanden waren 19 Jahre alt, der älteste Proband war 54 Jahre alt. Ein Proband war 30 und ein anderer 34 Jahre alt. Das Alter der restlichen Probanden lag ganz gleichmäßig verteilt zwischen 20 und 28 Jahren.

Zum Zeitpunkt der Teilnahme an der Studie hatten 5 Versuchspersonen ihr Studium abgeschlossen, ein Proband hat ein zweites Studium geführt und ein anderer promoviert. Die restlichen Versuchspersonen waren Studenten, davon haben 8 Probanden Informatik und 9 Softwaretechnik studiert. Insgesamt 18 Probanden trugen Sehhilfe, 13 davon trugen Brillen und 5 Kontaktlinsen.

4.5.2 Stimuli

Nach der Pilotstudie konnte die optimale Anzahl der zu präsentierenden Bäume endgültig festgelegt werden. In der eigentlichen Studie wurden 9 Varianten der Baumdarstellungen untersucht (Abb. 17, Abb. 18 und Abb. 19):

- Radiale Darstellung

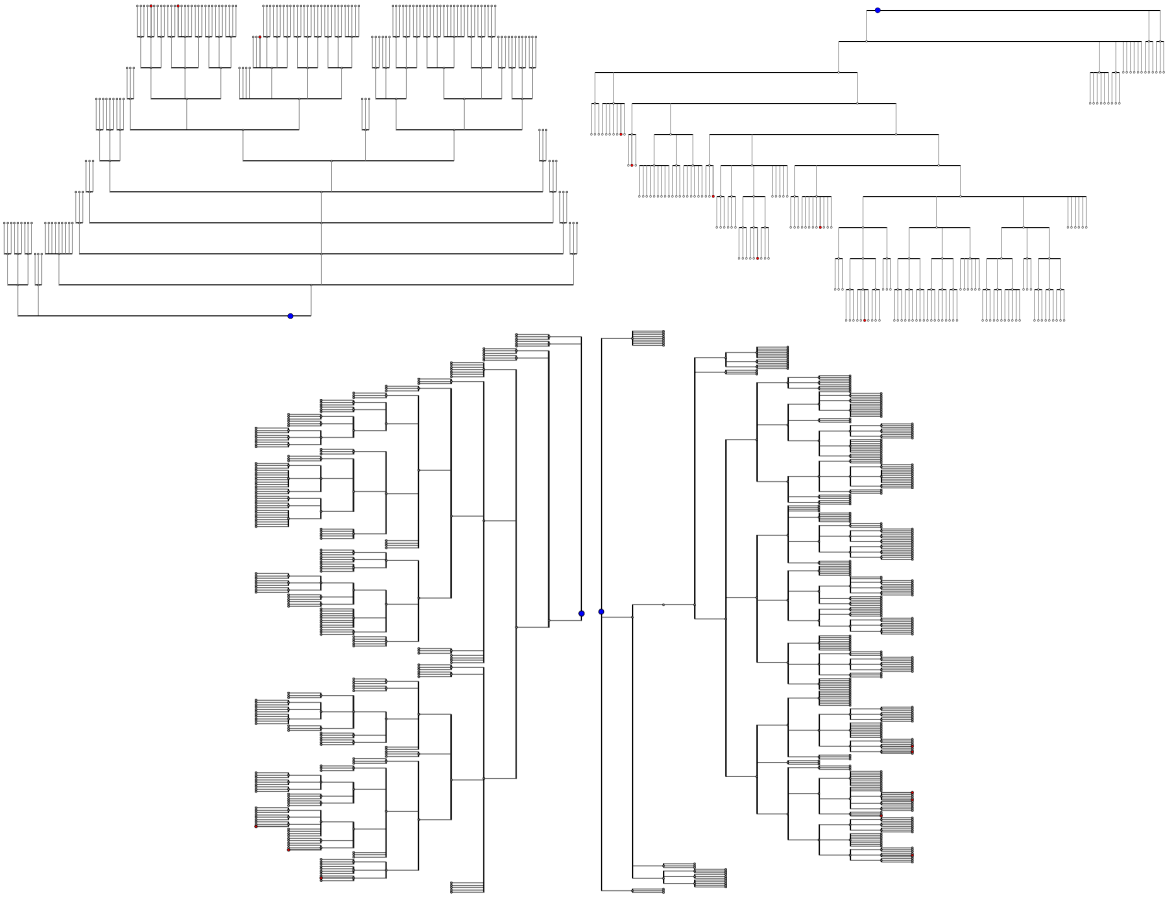

Abbildung 18: Beispiele der vier Varianten für die orthogonale Darstellungsart

- Traditionelle Darstellung (mit Wurzel oben, unten, links und rechts)
- Orthogonale Darstellung (mit Wurzel oben, unten, links und rechts)

Von jeder Darstellungsart wurden 6 zufällig generierte Bäume genommen. Diese Darstellungen wurden entsprechend der Art (traditionell, orthogonal und radial) in drei Blöcke aufgeteilt. Die Reihenfolge der Präsentation inmitten von jedem Block war zufällig, d.h. ein traditioneller Baum mit der Wurzel oben/unten konnte beispielsweise vor oder auch nach einem traditionellen Baum mit der Wurzel links/rechts eingeblendet werden.

Ein zusätzlicher Testblock bestand aus drei Bäumen (Abb. 20). Jeder Baum wurde für 30 Sekunden eingeblendet und der Proband sollte freie Kommentare zu der konkreten Baumdarstellung geben. Dieser Testblock wurde für jeden Teilnehmer als der letzte in dem ganzen Versuchsdurchlauf präsentiert.

4.5.3 Experimentumgebung

Die Aufnahmen der Blickbewegungen der Probanden wurden mit dem Eye-Tracker-System Tobii T60 XL durchgeführt. Die Bildschirmauflösung betrug 1920 x 1200 Pixel. Damit bei der Analyse

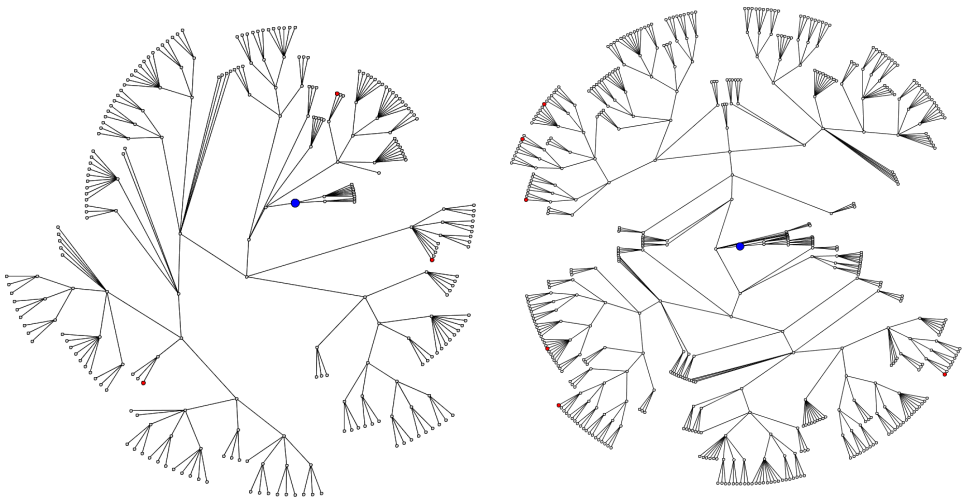

Abbildung 19: Beispiele für die radiale Darstellungsart

der möglichen Lösungsstrategien die feineren Zwischenschritte im Blickverhalten der Probanden sichtbar sind, wurde die ursprüngliche Einstellung des Eye-Trackers im „ClearView Fixation Filter“ übernommen. Sie betrug 10 Pixel für den Umkreis und 30 ms für die Dauer einer Fixation.

Die Fenster waren während des ganzen Studienablaufs verschlossen und verdunkelt, damit Geräusche und Bewegungen (Regen, Schnee, Wind) nicht zur Störvariablen werden. Der Raum war künstlich beleuchtet und enthielt nur die minimal notwendige Anzahl von Gegenständen. Vor allem wurde ständig kontrolliert, dass sich in der Nähe vom Eye-Tracker-Gerät keine überflüssige Sachen befinden. Vor Beginn jedes Testdurchlaufs wurden die Teilnehmer gebeten, ihre Mobiltelefone auszuschalten oder diese lautlos zu stellen.

4.5.4 Studienablauf

Der Ablauf der Studie hat je nach dem, wie schnell der konkrete Proband bei der Lösung der Aufgaben war und ob er die Pausen benötigt hat, zwischen 41 und 76 Minuten gedauert. Zuerst haben die Probanden die Aufklärung zum Ablauf der Studie durchgelesen und diese unterschrieben. Danach haben sie den Fragebogen über persönliche Daten (Geschlecht, Alter, Studiengang usw.) ausgefüllt. Als nächstes wurde der Sehtest und der Test auf Farbenblindheit durchgeführt.

Dann haben die Probanden das Tutorial durchgelesen. Gleich danach mussten sie die Blätter, die inneren Knoten und die Wurzel von den drei auf Papier ausgedruckten Beispielbäumen zeigen und auf drei weiteren Bäumen die eigentliche Testfrage beantworten. Informatik- und Softwaretechnikstudenten mussten die Testfrage beantworten, ohne vorher Wurzel, Blätter oder innere Knoten zu zeigen. Bei Studenten anderer Studiengänge hat nach dem Durchlesen des Tutorials zusätzlich eine ungefähr gleich verlaufende Beantwortung der Probandenfragen und eine Besprechung der Begriffe stattgefunden.

Sobald es klar wurde, dass die Teilnehmer sowohl die Begriffe als auch die Aufgabenstellung richtig verstanden haben, wurde ein mündlicher Überblick über den eigentlichen Testablauf ge-

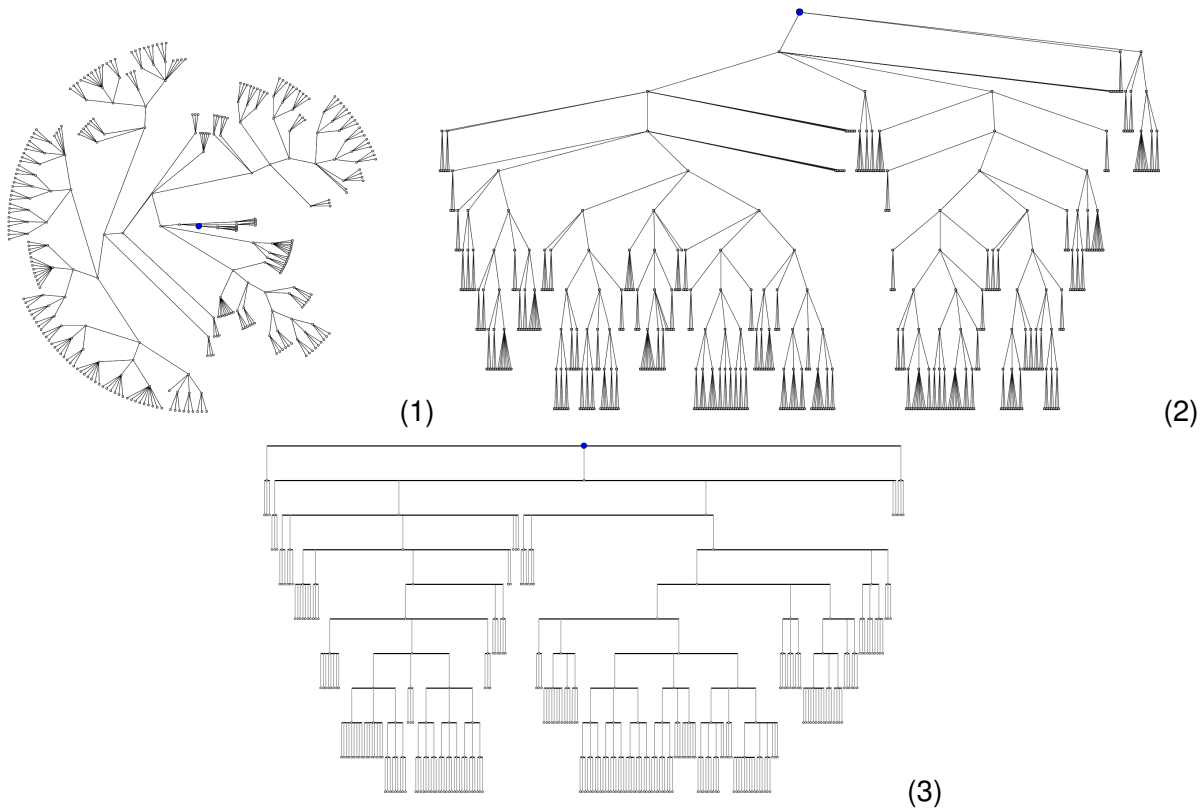

Abbildung 20: Stimuli für den Block mit offenen Fragen (die Reihenfolge wie im Testdurchlauf)

geben. Dieser Überblick bestand aus vorher festgelegten Sätzen, die in derselben Reihenfolge gesagt wurden.

Nach dem Überblick wurde der Probedurchlauf durchgeführt, der aus neun Bäumen (jeweils einem von jeder Darstellungsart) bestand. Die Baumtypen wurden in der Reihenfolge willkürlich gemischt und diese Reihenfolge blieb für alle Probanden gleich.

Als nächster Schritt wurde die eigentliche Studie durchgeführt. Dieser Abschnitt hat, je nach dem wie schnell ein Proband vorankam und ob er die Pausen benötigt hat, zwischen ca. 15 und ca. 45 Minuten gedauert.

Danach haben die Probanden einen Fragebogen mit Fragen über Baumdarstellungen ausgefüllt und organisatorische Fragen beantwortet. Sie wurden darüber befragt, wie sie auf die Studie aufmerksam geworden sind und ob sie noch Bemerkungen zu den Darstellungsarten oder dem Studienablauf haben.

Am Ende der Studie mussten sie die Liste für die Geldausgabe unterschreiben. Bei Interesse an der Teilnahme in weiteren Studien hatten sie die Möglichkeit, ihre E-Mail-Adresse in einer Adressenliste zu hinterlassen. Die meisten haben die Möglichkeit auch in Anspruch genommen.

Es gab in der Studie ein Ziel, die gleiche Anzahl der Aufnahmen für jede Permutation (beispielsweise 6 Aufnahmen für jede der 6 möglichen Permutationen) zu bekommen. Da in diesem Fall jeder Block genau gleich oft in der Studie auf der ersten, zweiten bzw. dritten Stelle vorkommt, sollten auf diese Weise Lern- und Reihenfolgeneffekte verhindert werden. Das Vorhaben konn-

te realisiert werden. Zwei zusätzliche Aufnahmen wurden als Ersatz für fehlerhafte Datensätze erstellt.

Da es im Test keine Lösungen gab, die sich ein Teilnehmer merken konnte, bestand auch keine Gefahr, dass ein Teilnehmer die nachfolgenden stark beeinflussen kann, indem er die möglichen Lösungen der Aufgaben verrät.

4.5.5 Ergebnisse

Es wurden insgesamt 4 x 38 Aufnahmen erzeugt (38 Probanden und 4 Testblöcke). Bei einem Proband trat das oben beschriebene Problem auf: Aus irgendeinem Grund hat der Eye-Tracker die Aufnahme kurz abgebrochen und dann eine zweite gestartet.

So ergaben sich für einen traditionellen Baum zwei Aufnahmen. Die Länge der ersten war ca. 5 Sekunden. Die Heat Map für diese Aufnahme enthielt keinen Mausklick. Die Länge der zweiten war ca. 20 Sekunden und der Mausklick war vorhanden. Es war unklar, ob die Lösungsdauer aus der Summe der beiden Längen errechnet werden sollte. Deswegen musste der gesamte Datensatz von diesem Probanden verworfen werden und durfte in der Auswertung nicht benutzt werden.

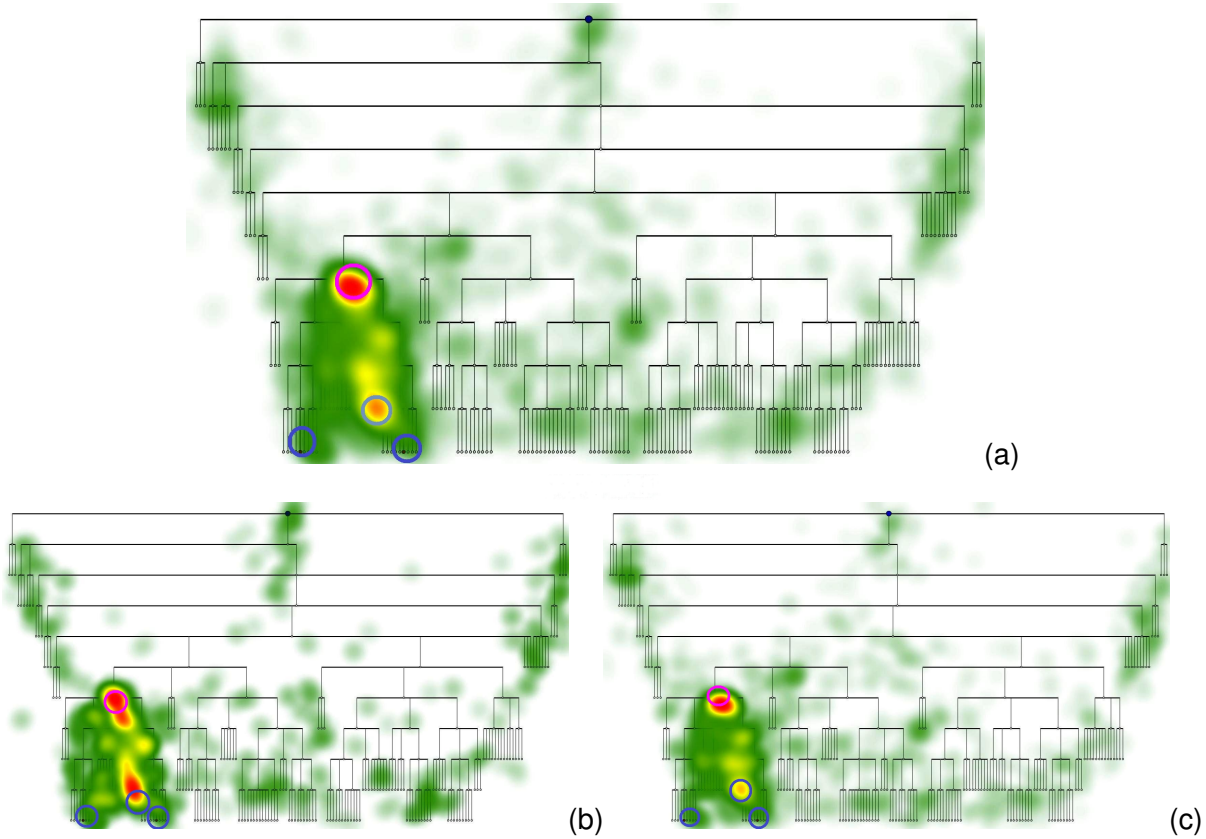

Abbildung 21: Heat Maps einer orthogonalen Darstellung; (a) - gesamt, (b) - Frauen, (c) - Männer

Der andere unbrauchbare Datensatz entstand beim ersten Testdurchlauf. Die Ursache dafür war eine falsche Einstellung für einige traditionelle Baumdarstellungen. Die Option “Show Mouse Cursor” war nicht aktiviert, so dass der Proband keine Möglichkeit hatte, die Lösung mit

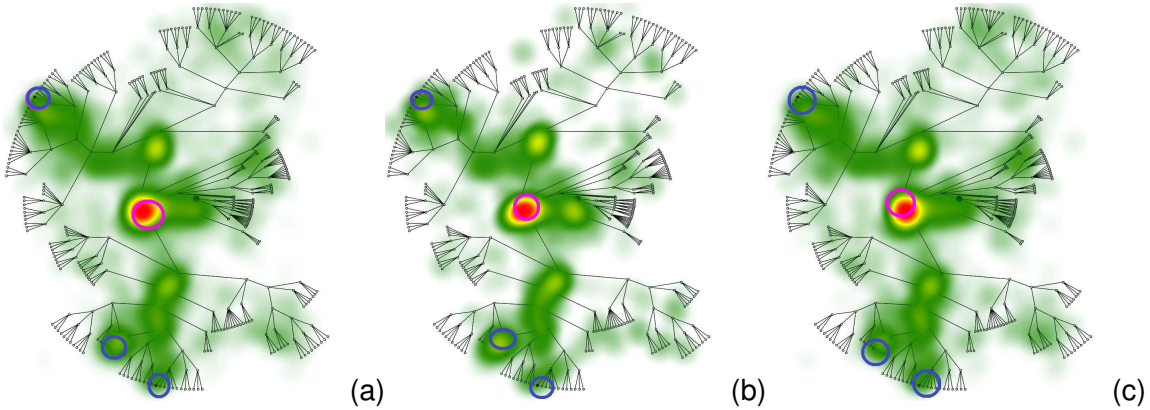

Abbildung 22: Heat Maps einer radialen Darstellung; (a) - gesamt, (b) - Frauen, (c) - Männer

der Maus anzuklicken. Der Fehler konnte schnell behoben werden und trat bei den weiteren Testdurchläufen nicht mehr auf.

Da die beiden genannten Fehler früh entdeckt wurden, konnten bei der Durchführung der 37. und der 38. Aufnahme die passenden Permutationen der Blöcke verwendet werden.

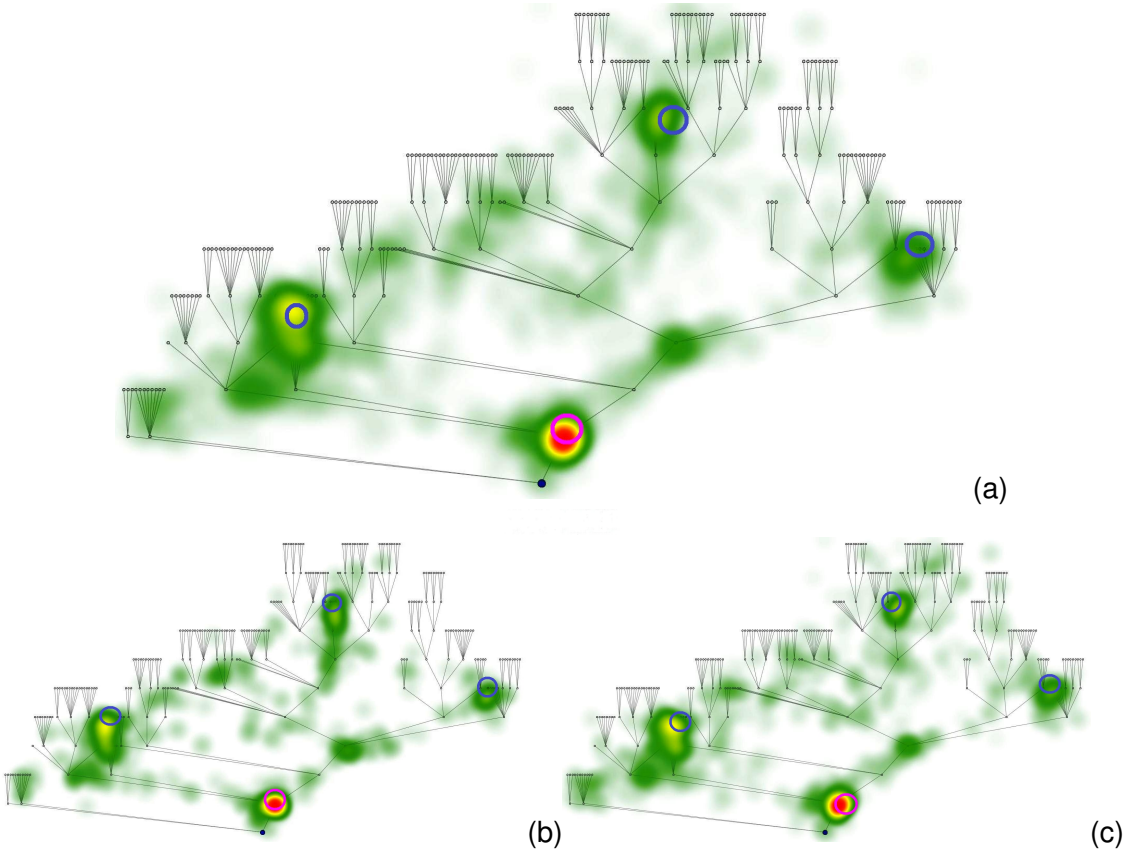

Abbildung 23: Heat Maps einer traditionellen Darstellung; (a) - gesamt, (b) - Frauen, (c) - Männer

4.6 Phase der Auswertung

Die genauere Beschreibung der Auswertungsphase sowie der Abbildungen 21 bis 26 befindet sich im Kapitel 5.

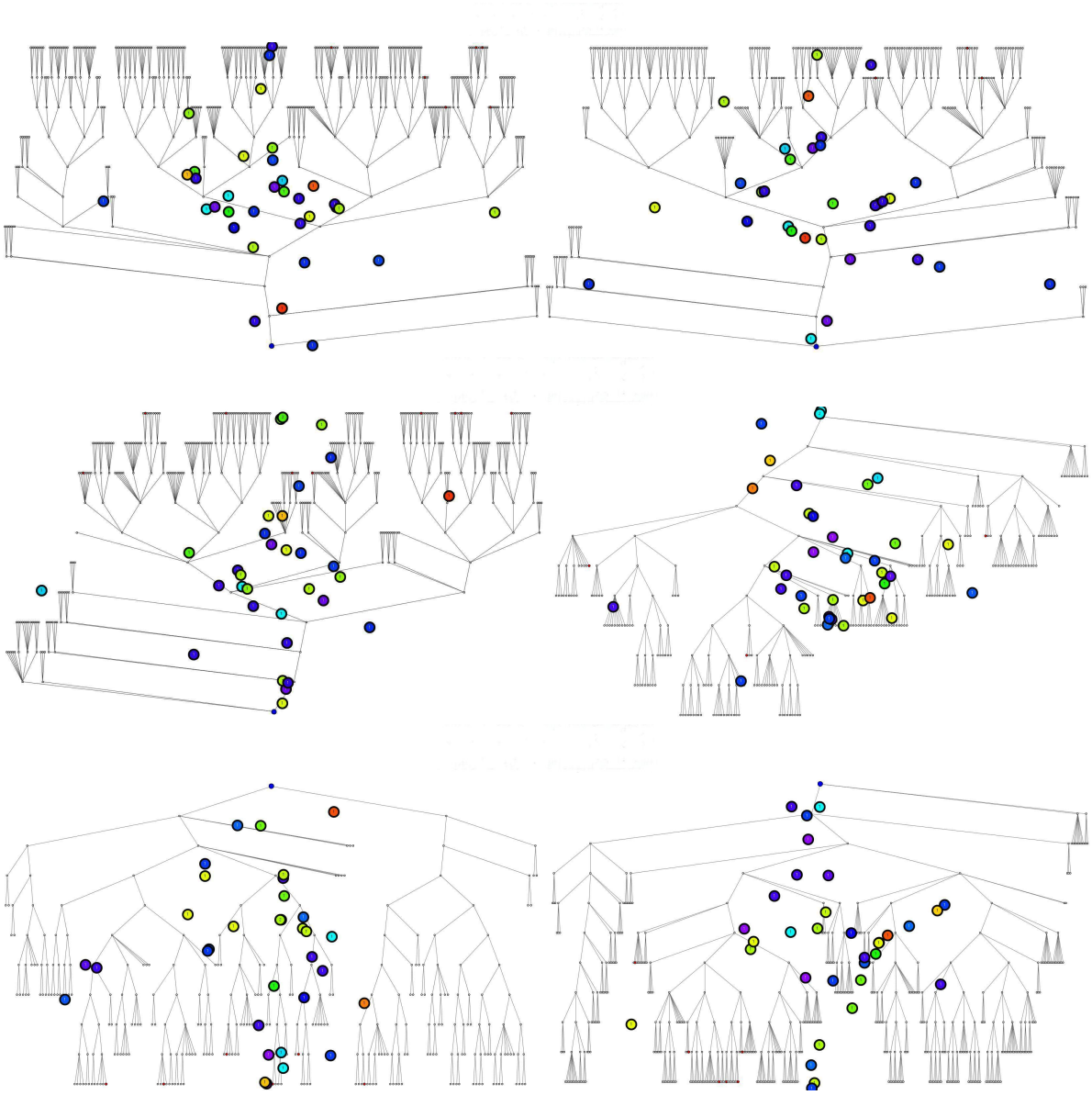

Abbildung 24: Der erste Blick bei der traditionellen Dartellung (Wurzel oben und unten)

4.7 Phase der Publikation

Die Vergleichsstudie und ihre Ergebnisse wurden dokumentiert und mussten im Rahmen dieser Diplomarbeit veröffentlicht werden. Um den Studienaufbau kompakt und übersichtlich darzustellen, wurden die Strukturierung des Abschnitts 3.1 und die sich im Kapitel 3 befindenden Fragenstellungen verwendet.

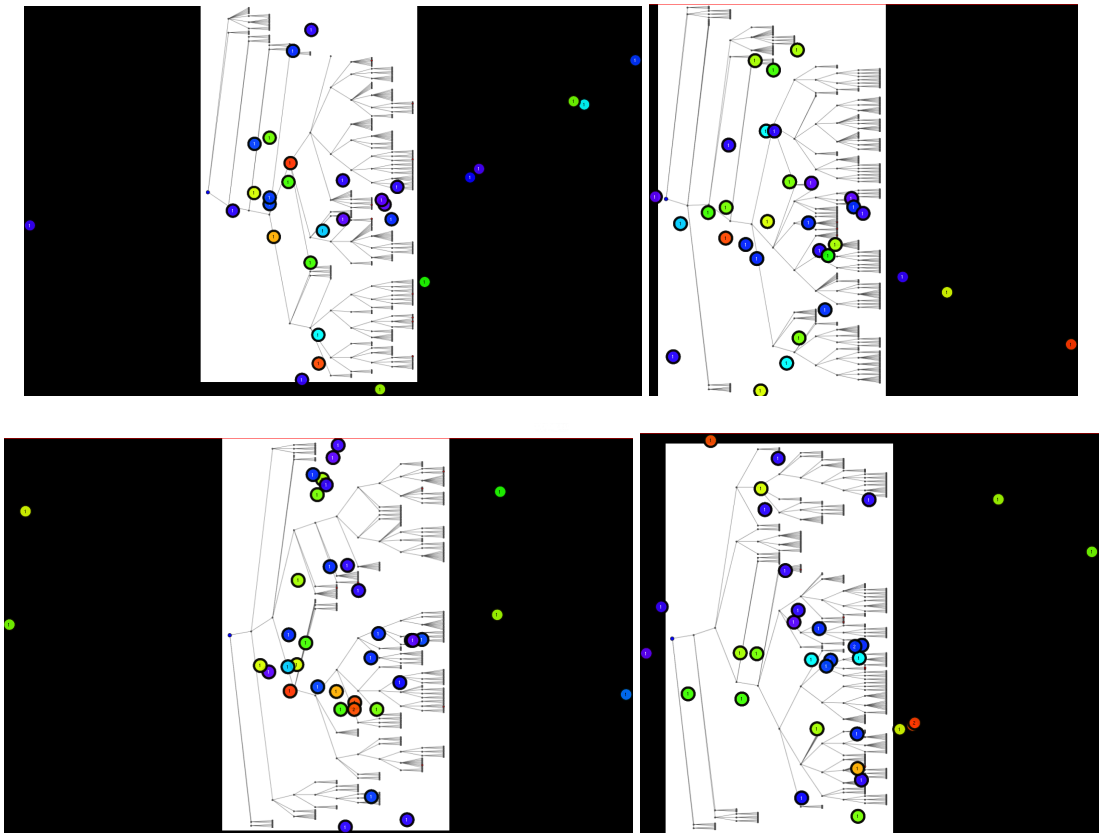

Abbildung 25: Der erste Blick bei der traditionellen Dartellung (Wurzel links)

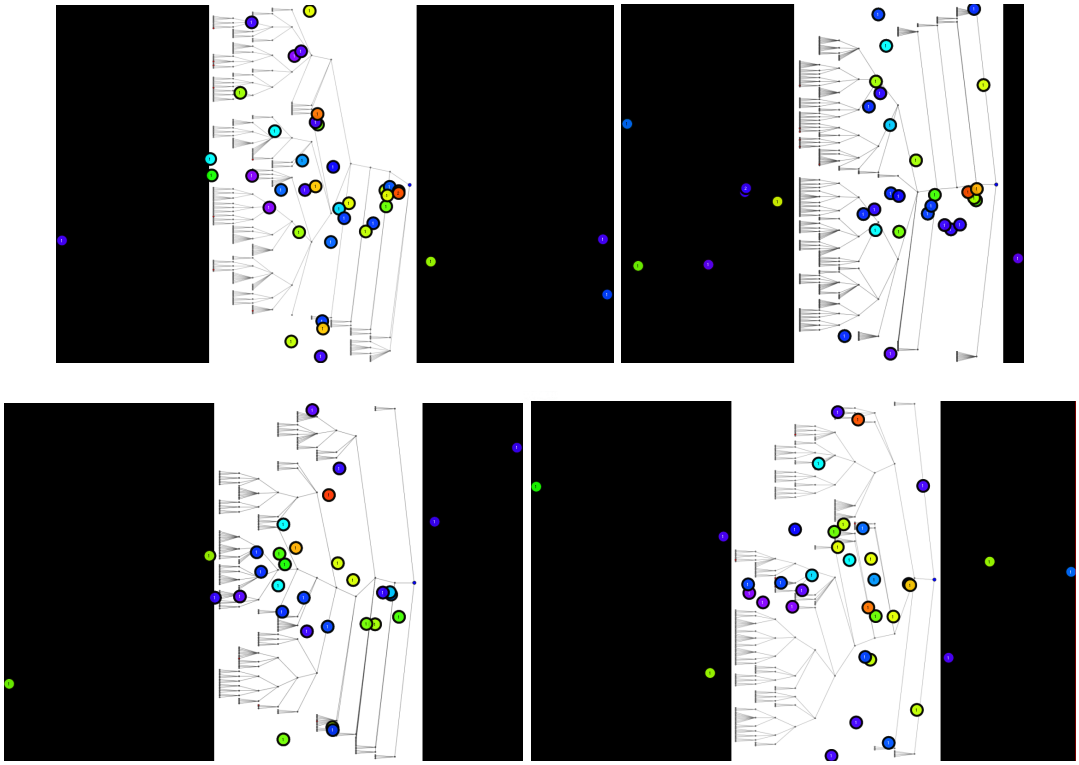

Abbildung 26: Der erste Blick bei der traditionellen Dartellung (Wurzel rechts)

5 Auswertung

Gleich nachdem die Entscheidung getroffen wurde, in der Studie traditionelle, orthogonale und radiale Baumdarstellungen zu vergleichen, war es möglich, eine Hypothese für die zu untersuchende Frage aufzustellen. Es wurde vermutet, dass die Suche in einem traditionellen Baum am schnellsten und in einem radialen Baum am langsamsten sein wird. Die Ergebnisse der statistischen Auswertung sind im Abschnitt 5.3 präsentiert.

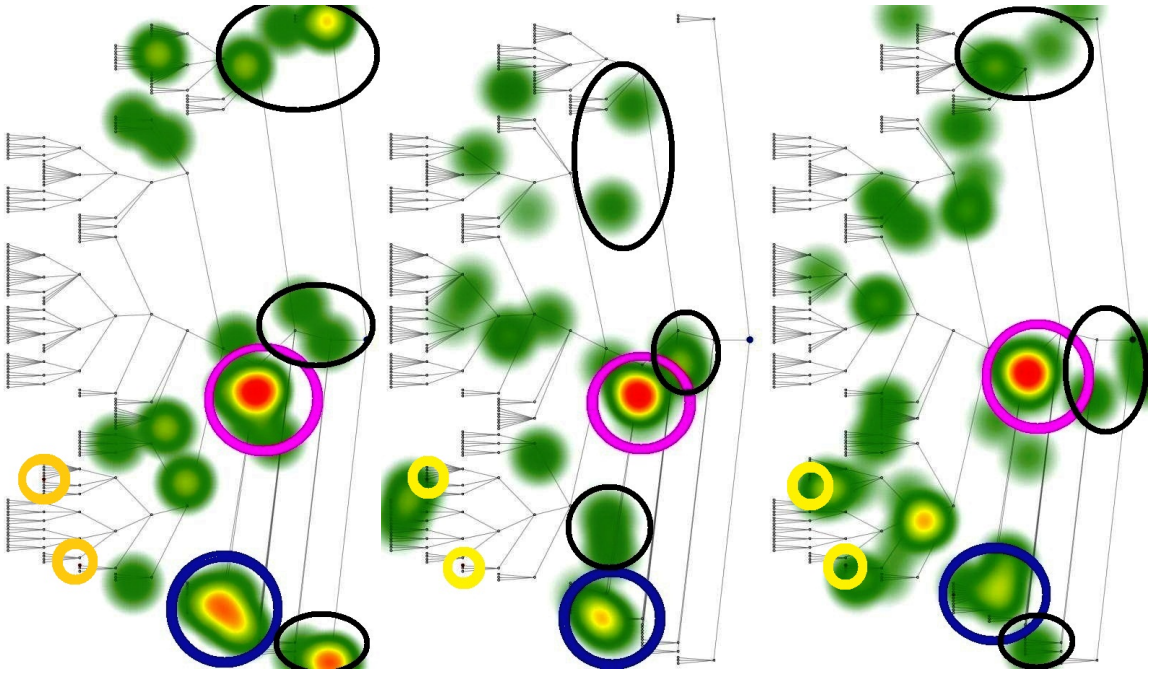

Abbildung 27: Strategisches Vorgehen bei der Suche nach der Lösung; Heat Maps von den einzelnen Probanden

5.1 Heat Maps

Auf den Heat Maps sind die Bereiche sichtbar, die von den Testpersonen besonders oft und intensiv betrachtet wurden. So läßt sich das Blickverhalten analysieren und die möglichen Strategien erkennen, die die Probanden bei der Lösung der Aufgaben entwickeln. Ähnlich wie bei den farbkodierten Höhenschichten in der Topografiedarstellung sind hier die am meisten betrachteten Abschnitte des Bildes rot gefärbt. Die Abschnitte, die überhaupt nicht angeschaut wurden, bleiben bei den Heat Maps von einzelnen Probanden ohne Färbung (Abb. 27). Es ist möglich, dass in der Heat Map von mehreren Personen bestimmte Bereiche im Vergleich zu den einzelnen Heat Maps weißer bzw. grüner werden und die anderen gelber bzw. roter (Abb. 21, Abb. 22, Abb. 23). Das ist damit verbunden, dass die Werte der einzelnen Personen einander ausgleichen und in der Summe die neue Farbverteilung ergeben.

Die in diesem Kapitel präsentierten Heat Maps wurden anhand der Anzahl der Fixationen erstellt. Tobii T60 XL bietet noch zwei alternative Möglichkeiten zur Erstellung von Heat Maps.

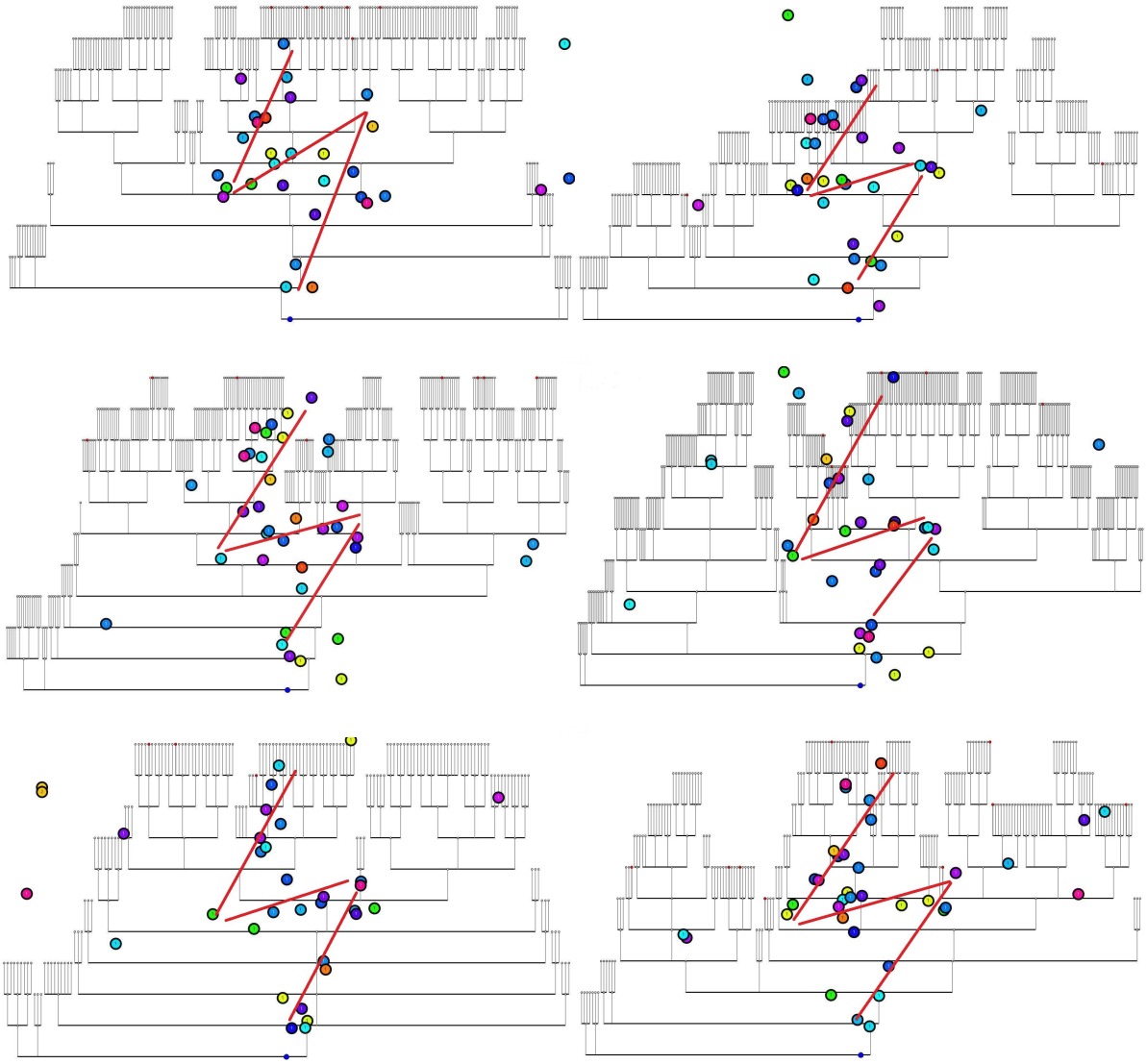

Abbildung 28: S-Muster bei orthogonalen Darstellungen mit der Wurzel unten

Es ist möglich, die Heat Maps abhängig von absoluter oder relativer Dauer der Fixationen zu erhalten.

Die Auswertung der Heat Maps hat gezeigt, dass die Probanden am längsten die markierten Blätter (blau) und die Lösungsknoten (rosa) anschauen (Abb. 21, Abb. 22, Abb. 23). Mehrere Heat Maps gaben einen Hinweis auf eine mögliche Strategie beim Lösen der Aufgabe. Sie bestand darin, möglichst früh das markierte Blatt zu identifizieren, das am höchsten im Baum liegt. Wenn solch ein Blatt gefunden wird, werden die tieferliegenden Abschnitte kaum noch angeschaut und die Lösung in derselben oder in den darüberliegenden Hierarchien gesucht (Abb. 27). In der Abbildung sind mit blau das höchstliegende markierte Blatt, mit gelb die tiefliegende markierte Blätter, mit rosa der Lösungsknoten und mit schwarz die Suchbereiche in oberen Hierarchien gekennzeichnet. Es fällt auf, dass die unterste Blattebene sowie die am tiefsten liegenden markierten Blätter kaum noch angeschaut wurden.

Es wurden keine auffälligen Unterschiede im Suchvorgang zwischen den Geschlechtern fest-

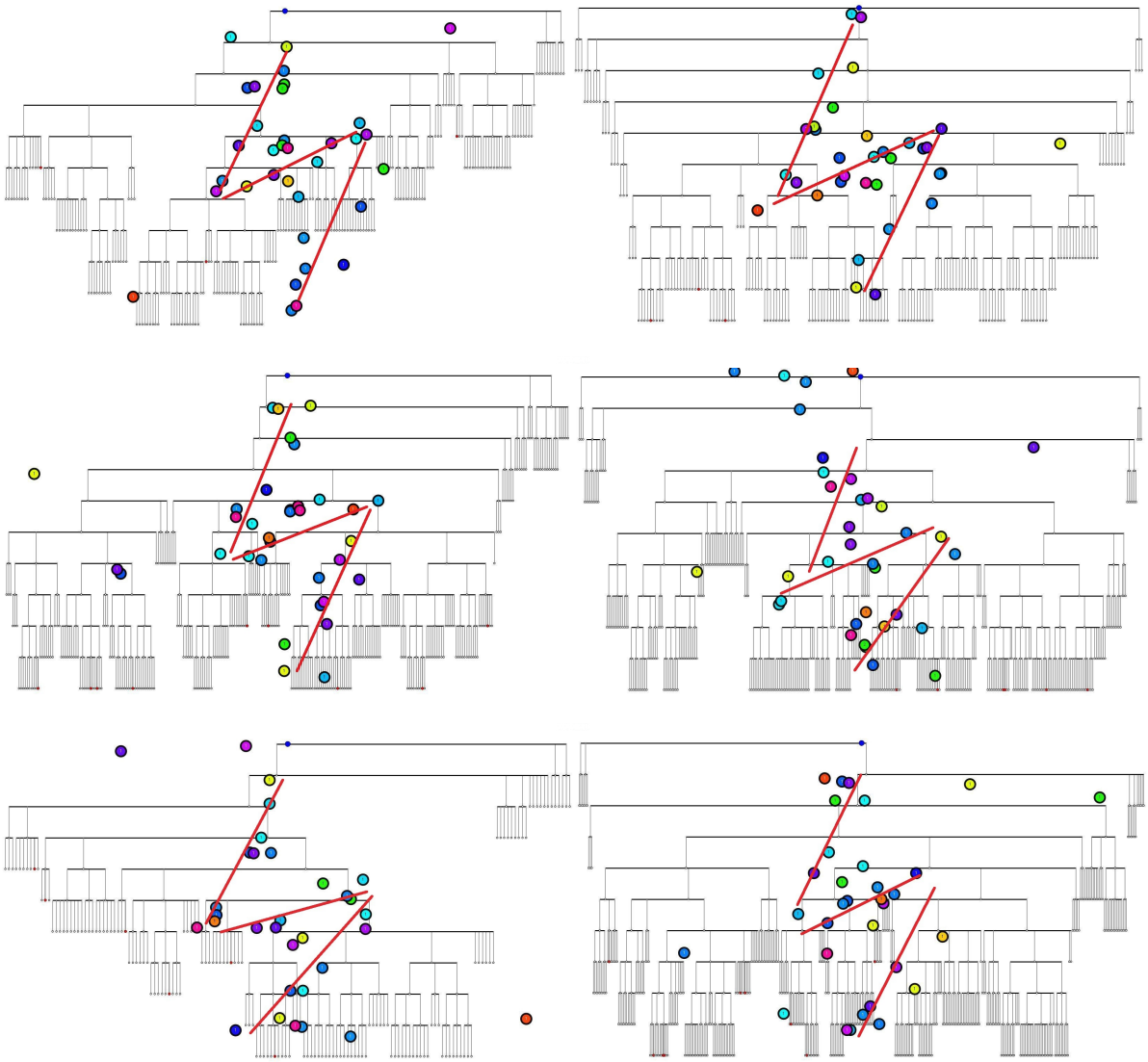

Abbildung 29: S-Muster bei orthogonalen Darstellungen mit der Wurzel oben

gestellt. Die Heat Map der Frauen unterscheidet sich nur gering von der Heat Map der Männer (Abb. 21, Abb. 22, Abb. 23). Dieser Unterschied ist auf die Anzahl der Probanden zurückzuführen (9 und 27 ohne fehlerhafte Datensätze), deren einzelne Heat Maps in eine Heat Map Darstellung zusammengefasst wurden.

5.2 Gaze Plots

Auf Gaze Plots lassen sich die Blickbewegungen eines einzelnen Probanden schrittweise untersuchen. Gaze Plots werden als Sakkaden (Linien) und Fixationen (Kreise) dargestellt. Die Nummerierung der Kreise ergibt eine hervorragende Möglichkeit, die einzelnen Blickpunkte der Betrachter der Reihe nach anzuschauen und zu analysieren.

Mit Hilfe von Gaze Plots sollte festgestellt werden, wohin die Probanden zuerst schauen.

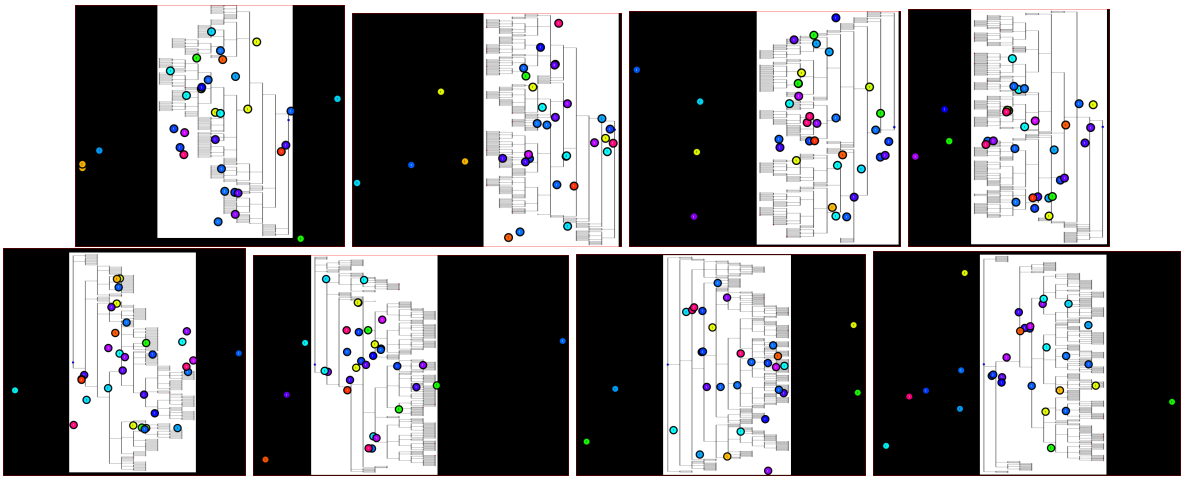

Abbildung 30: Der erste Blick bei der orthogonalen Dartellung (Wurzel links und rechts)

5.2.1 Gaze Plots von traditionellen Darstellungen

Bei traditionellen Bäumen (Wurzel oben und unten) schauen die Probanden zuerst ungefähr in die Mitte des Bildes und zwar in die Umgebung des sich zentral befindenden Pfades. Bei jedem Baum verschiebt sich die „Blickwolke“ entsprechend der Lage des mittleren Pfades (Abb. 24). Ähnliche Tendenz ist auch bei traditionellen Bäumen mit der Wurzel links bzw. rechts erkennbar (Abb. 25, Abb. 26).

5.2.2 Gaze Plots von orthogonalen Darstellungen

Bei orthogonalen Darstellungen mit der Wurzel oben bzw. unten ist ein interessantes Muster erkennbar. Es liegt in der Mitte der Darstellungen und hat eine Ähnlichkeit mit dem Buchstaben “S”. Dieser Muster ist bei allen 12 Bäumen gut sichtbar (Abb. 28, Abb. 29).

Die mittleren Pfade sind bei orthogonalen Darstellungen zwar nicht so kompakt, es ist aber erkennbar, dass der obere bzw. untere Abschnitt von “S” in der Nähe oder sogar direkt über dem mittleren Pfad liegt und zwar über seinem sich an der Wurzel befindendem Anteil.

Für orthogonale Darstellungen mit der Wurzel links bzw. rechts konnten keine Zusammenhänge mit der Form oder Lage des Baums festgestellt werden (Abb. 30). Die Verteilung der Blickpunkte ist auf dem Baum ganz gleichmäßig. Einmal liegen die Blickpunkte näher zur letzten Blattebene, ein anderes Mal näher zu den oberen Hierarchien. Es ist keine Systematik für ihre Platzierung in der Darstellungsart erkennbar. Es lässt sich aber vermuten, dass auch hier die Tendenz, den mittleren Pfad zuerst anzuschauen, wahrscheinlich ist, gäbe es nicht drei Schwierigkeiten:

- Die Darstellungen sind zu klein (ca. 65% kleiner als die mit der Wurzel oben und unten)
- Die mittleren Pfade bei orthogonalen Bäumen sind oft nicht kompakt
- Der Baum “liegt” auf der Seite

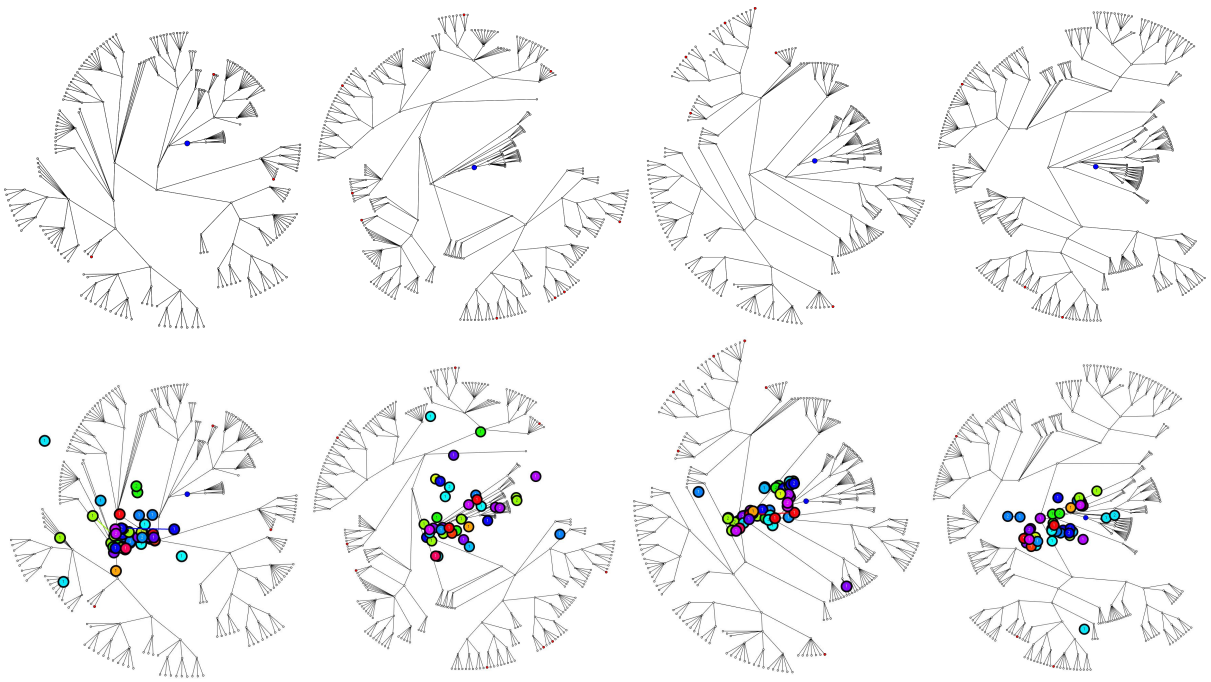

Abbildung 31: Der erste Blickpunkt bei den unsymmetrischen radialen Darstellungen

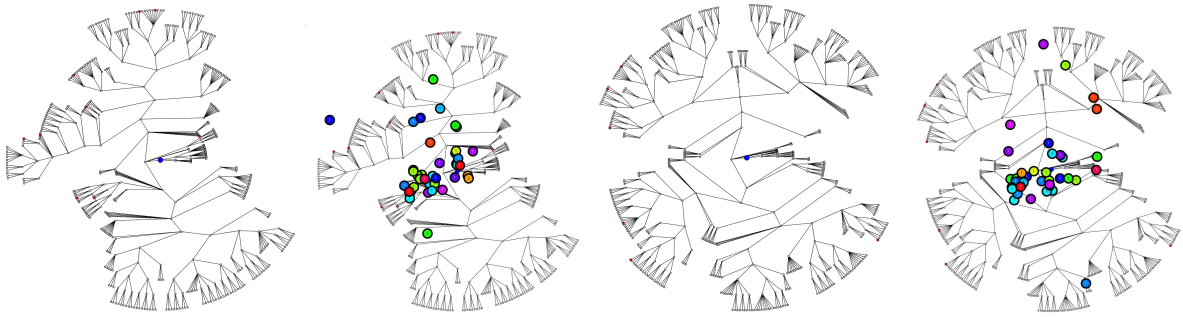

Abbildung 32: Der erste Blickpunkt bei den symmetrischen radialen Darstellungen

5.2.3 Gaze Plots von radialen Darstellungen

Für radiale Bäume konnte ein Zusammenhang zwischen Symmetrie der Darstellung und dem ersten Blick festgestellt werden.

Bei unsymmetrischen Bäumen haben die Probanden zuerst auf die inneren Knoten geschaut, die oben in der Hierarchie liegen und die mögliche Lösung sein könnten (Abb. 31). Eventuell haben die Teilnehmer an diesen Stellen die Wurzel erwartet. Eine alternative Vermutung ist interessanter und stellt den direkten Zusammenhang zur Systemantik des Blickverhaltens bei traditionellen und orthogonalen Darstellungen her.

In allen vier unsymmetrischen Darstellungen zieht sich die “Blickwolke” eindeutig von der Mitte des Bildes in die Richtung zum linken unteren Eck. Im Unterschied zu den beiden ganz symmetrischen Darstellungen ist bei den unsymmetrischen Bäumen der mittlere Pfad gut erkennbar und liegt direkt unter der “Blickwolke”. Diese endet genau dort, wo der Pfad sich besonders stark in entgegengesetzte Richtungen verzweigt und zu den größeren Teilbäumen führt.

Da bei den symmetrischeren Bäumen zwei “Hauptpfade” existieren (Abb. 32), kann diese Tatsache erklären, warum die ersten Blicke der Pobanden sich eher in der Mitte der Darstellung häufen. Wobei auch bei diesen beiden Bäumen eine gewisse Tendenz zur Verfolgung der ausgeprägteren Pfade (nach oben) erkennbar ist.

5.3 Statistische Auswertung

Da der Eye-Tracker keine Funktion zum automatischen Export der Zeiten anbietet, wurden die Antwortzeiten in eine Excel-Tabelle manuell übertragen. Sobald eine Menge aus 4-8 Datensätzen vorhanden war, wurden die Zeiten in die Tabelle übertragen. So wurde die Eingabe in einigen Durchläufen erledigt. Diese Aufteilung war zum einen dafür gedacht, die Zahl der Eingabefehler zu minimieren, die durch Monotonie und Konzentrationsverlust entstehen. Zum anderen waren die Daten gegen einen möglichen Ausfall des Eye-Trackers auf diese Weise gesichert. Zusätzlich wurden sie dadurch indirekt auf mögliche Messfehler und Inkonsistenzen frühzeitig geprüft. Hätte diese Prüfung viele fehlerhafte Datensätze aufgedeckt, wären rechtzeitig Maßnahmen getroffen worden, um noch mehr Probanden für die Studie zu bekommen.

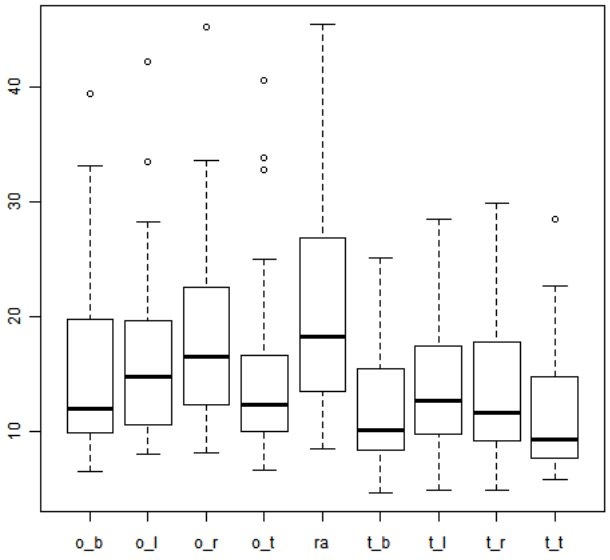

Abbildung 33: Die Boxplots mit den Mittelwerten und den Ausreißern. Von links nach rechts: orthogonale (Wurzel unten, links, rechts und oben), radiale und traditionelle Baumdiagramme (Wurzel unten, links, rechts und oben)

Als die Tabelle mit den Daten vollständig war, wurde sie systematisch überprüft. Jeder Eintrag wurde mit der entsprechnden Zeitangabe im Eye-Tracker verglichen. Diese Prüfung hat 28 Tippfehler (ca. 1,5%) aufgedeckt. Für die zweite Prüfung wurden einige Abschnitte der Tabelle zufällig ausgewählt. Aus 250 Einträgen war nur einer falsch (0,4%). Alle Konsistenz- und Glaubwürdigkeitsfragen aus dem Abschnitt 3.2.2 wurden während der Überprüfung geklärt. Sie haben einen konsistenten und glaubwürdigen Datensatz bestätigt. Die Ausreißer in der Boxplotdarstellung haben sich nicht als Mess- oder Tippfehler erwiesen (Abb. 33). Da ca. 85% aller Tippfehler in den beiden genannten Prüfschritten im Intervall von 0,01 bis 1 Sekunde lag, wurde keine weitere Überprüfung mehr durchgeführt.

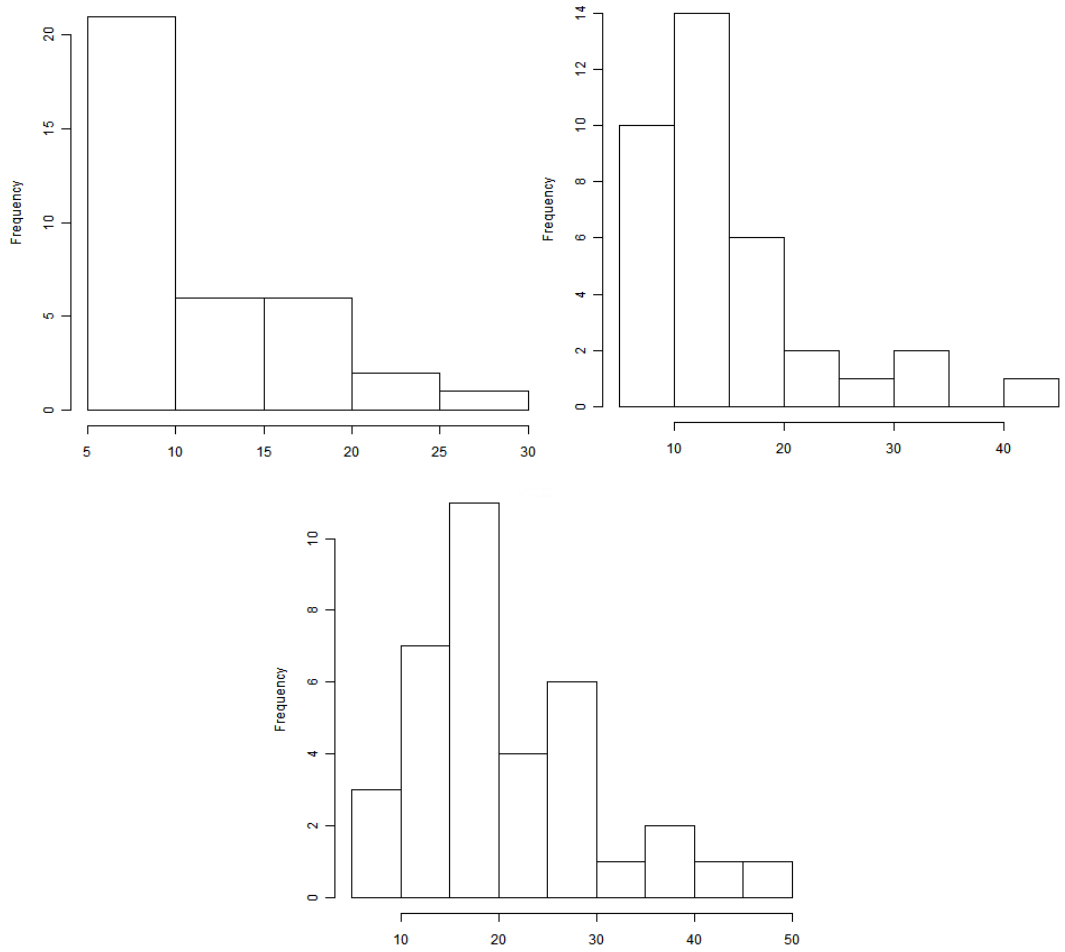

Abbildung 34: Verteilung der Daten vor der Transformation (traditionell, orthogonal, radial); Sekunden auf der x-Achse, Anzahl von Probanden auf der y-Achse

Die statistische Auswertung der Daten wurde unter Verwendung von R durchgeführt [21]. Da für den ersten Vergleich (siehe Abschnitt 4.2) drei Stichproben untersucht werden sollten, war dafür von vornherein die Analyse mit ANOVA eingeplant. Damit ANOVA angewendet werden darf, müssen die zu untersuchenden Daten Varianzhomogenität (die Varianz der Daten ist überall annähernd gleich) aufweisen und nicht stark von der Normalverteilung abweichen [4]. Für die Auswertung wurden die Antwortzeiten der Probanden über alle 6 Darstellungen gemittelt, d. h. die Antwortzeit für die orthogonale Darstellung mit der Wurzel rechts ergab sich als Durchschnitt aus den Anwortzeiten aller 6 othogonalen Bäumen mit der Wurzel rechts. So konnte verhindert werden, dass mögliche Unterschiede durch die konkreten Baumformen verursacht werden. Die Boxplots stellen die Mittelwerte für die neun Darstellungsarten über alle 36 Aufnahmen dar (Abb. 33). Die gesammelten Daten wichen zunächst von einer Normalverteilung ab (Abb. 34), deswegen wurden sie mit Hilfe des Logarithmus' transformiert (Abb. 35).

Die transformierten Daten wurden per Shapiro-Wilk-Test auf Normalität und per Bartlett-Test auf Varianzhomogenität geprüft. Die Tests haben die beiden Eigenschaften nachgewiesen .

Die Analyse mit ANOVA hat gezeigt, dass signifikante Unterschiede zwischen den Baumdarstellungen vorhanden sind. Um festzustellen, zwischen welchen konkreten Darstellungen der

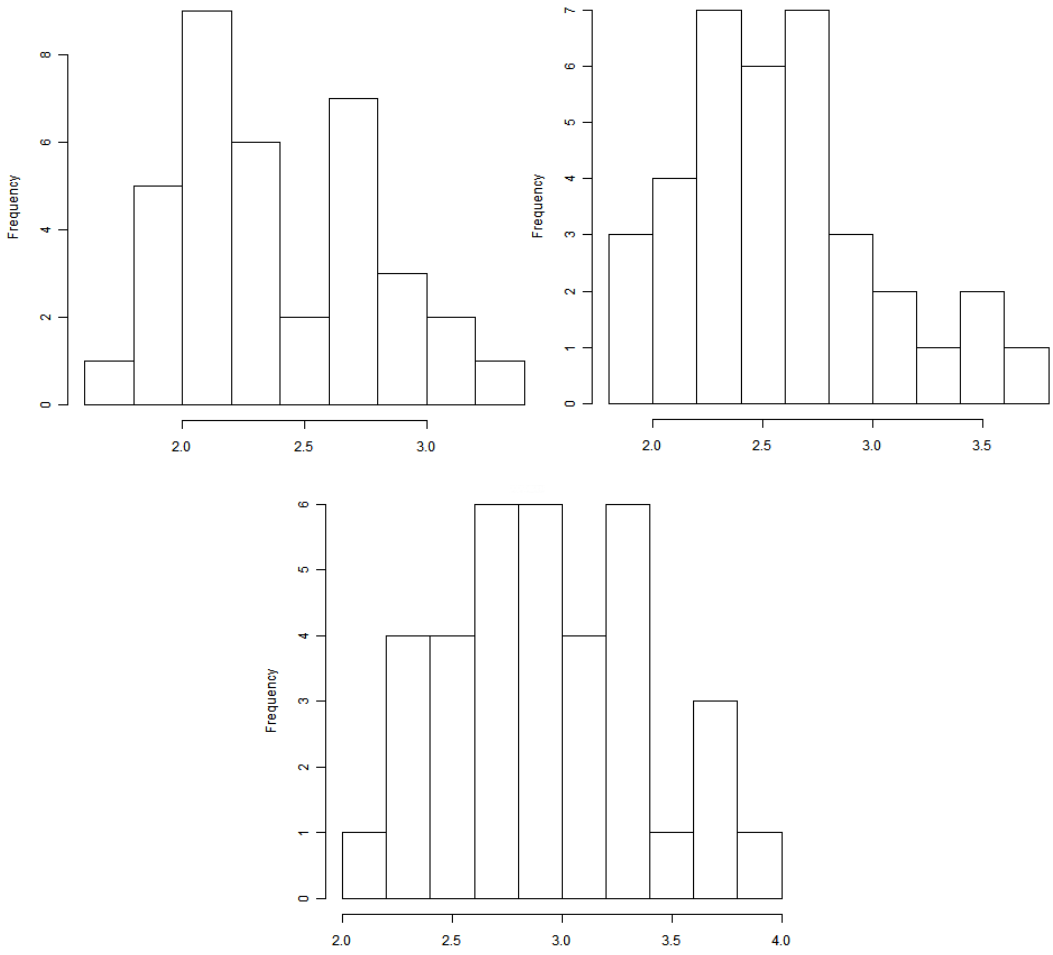

Abbildung 35: Verteilung der Daten nach Logarithmustransformation (traditionell, orthogonal, radial); ln(t) auf der x-Achse, Anzahl von Probanden auf der y-Achse

Unterschied vorliegt, wurde der Tukey-HSD-Test durchgeführt. Der Test hat gezeigt, dass zwischen traditionellen und radialen Baumdarstellungen (p=6×10^{-7}) und zwischen orthogonalen und radialen Darstellungen (p=$26,144 \times 10^{-4}$) signifikante Unterschiede vorliegen. Zwischen traditionellen und orthogonalen Darstellungen ergab sich ein nahezu signifikanter Unterschied (p=$81,8088 \times 10^{-3}$).

Für den zweiten Vergleich (siehe Abschnitt 4.2) wurde der Zweistichproben t-Test von Welch durchgeführt. Der paarweise Vergleich der Mittelwerte hat gezeigt, dass zwischen keinem von vier Paaren signifikante Unterschiede existieren.

6 Zusammenfassung

Der Vergleich von Visualisierungen ist ein sehr aktueller und besonders interessanter Forschungsbereich. Dabei ist das Interessante gleichzeitig auch das Herausfordernde und besteht darin, dass es im Visualisierungsbereich sowohl um den Computer als auch um den Menschen und zwar um sein visuelles Wahrnehmen geht.

Die Visualisierungskomponenten werden entwickelt, um den Benutzern die Arbeit mit den großen Datenmengen zu erleichtern. Die wichtige Frage besteht darin, ob sie tatsächlich und wie gut diesem Zweck dienen.

Damit diese Frage untersucht werden kann, ist eine sehr systematische Vorgehensweise gefordert. Diese Arbeit gibt einen Vorschlag für die Strukturierung und den Design einer Vergleichstudie im Visualisierungsbereich. Beim Entwurf und der Durchführung solch einer Studie sowie bei der Auswertung ihrer Ergebnisse sind sehr viele Aspekte und Fragen zu klären. Diese Aspekte und Fragen könnten in sieben Blöcke entsprechend der Phase der Studie gruppiert werden. Solch ein Gerüst hilft, planvoll und durchdacht vorzugehen sowie keine relevanten Facetten der Aufgabe außer Acht zu lassen.

Damit dieser Designvorschlag selbst auf mögliche Schwächen und Lücken überprüft werden konnte, wurde im Rahmen dieser Arbeit eine Vergleichstudie für Hierarchievisualisierungen durchgeführt. Es hat sich gezeigt, dass der Vorschlag eine sehr gute Stütze in der Arbeit ist und alle möglichen Seiten der Aufgabenstellung beleuchtet.

Die durchgeführte Studie hat nur einen sehr kleinen Ausschnitt aus der großen Palette der möglichen Fragestellungen im Visualisierungsbereich untersucht. Nichtsdestotrotz liefert sie einen wichtigen Beitrag für die empirische Forschung.

Durch die starke Abgrenzung des Problembereichs wurde es möglich, die Studie in der Form eines kontrollierten Experiments durchzuführen. Diese Form ist notwendig, um die individuellen Unterschiede der Testpersonen auszugleichen und wissenschaftlich fundierte Ergebnisse zu bekommen. Die Form des kontrollierten Experiments ist für alle Vergleichsstudien im Visualisierungsbereich zu empfehlen.

Der Fokus der Studie lag auf dem Vergleich von unterschiedlichen Darstellungsarten von Baumdiagrammen. Die herausragenden Ergebnisse bestehen darin, dass die signifikanten Unterschiede in der Antwortzeit bei der Lösung der Testaufgaben festgestellt wurden. Bei der Suche in den traditionellen Darstellungen sind die Probanden signifikant schneller als bei radialen. Die signifikanten Unterschiede wurden ebenso zwischen orthogonalen und radialen Darstellungen festgestellt.

Zusätzlich konnte eine für mehrere Probanden gemeinsame Lösungsstrategie festgestellt werden: Zum bestimmten Zeitpunkt lernen die Probanden, dass der Lösungsknoten nicht tiefer als der höchste markierte Knoten liegen kann. So schauen sie zuerst nach den Markierungen in den mittleren und höheren Hierarchien nach, um schneller die Aufgabe lösen zu können.

Abschließend lässt sich bemerken, dass die vorliegende Arbeit allen eine gute erste Orientierung im beschriebenen Bereich geben kann, die die ähnlichen Vergleichsstudien durchführen wollen.

Literatur

[1] C. Arnold. *Visualisierung in Information Retrieval*. VDM Verlag Dr. Müller, Saarbrücken, 2008.

[2] J. Bortz. *Statistik für Sozialwissenschaftler*. Springer, Berlin [u.a.], 1999.

[3] J. Bortz and N. Döring. *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Springer, Berlin [u.a.], 2006.

[4] F. Brosius. *SPSS 8 Professionelle Statistik unter Windows*. mitp, 1998.

[5] M. Burch, F. Bott, F. Beck, and S. Diehl. *Cartesian vs. Radial - A Comparative Evaluation of Two Visualization Tools - in Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, 2008.

[6] M. Burch, M. Raschke, and D. Weiskopf. *Indented Pixel Tree Plots*. ISVC (1) 2010, S. 338 - 349, 2010.

[7] W. Cleveland and R. McGill. *Dynamic Graphics for Statistics*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1988.

[8] R. Fukuda and H. Bubb. *Eye tracking study on Web-use: Comparison between younger and elderly users in case of search task with electronic timetable service*. PsychNology Journal, 2003 Volume 1, Number 3, S. 202 - 228, 2003.

[9] V. Gollücke. *Eye-Tracking - Grundlagen, Technologien und Anwendungsgebiete*. GRIN Verlag, 2009.

[10] L. Granka, T. Joachims, and G. Gay. *Eye-Tracking Analysis of User Behavior in WWW-Search*. Poster Abstract, Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR), S. 478 - 479, 2004.

[11] D. Holten and J. Wijk. *A User Study on Visualizing Directed Edges in Graphs*. CHI, Boston, 2009.

[12] D. Keim, F. Mannsmann, A. Stoffel, and H. Ziegler. *Visual Analytics. In: Encyclopedia of Database Systems*. Springer, 2009.

[13] D. Keim, S. North, C. Panse, and M Sips. *Visual Data Mining in Large Geo-Spatial Point Sets. In: IEEE Computer Graphics and Application. Nr. 12, 2004, S. 36 - 44.* Reading, Addison-Wesley, 2004.

[14] M. Kenner. *Einführung in die Statistik, Skript zur Vorlesung Forschungsmethoden der BWP", WS 2008/09*. Universität Stuttgart Institut für Erziehungswissenschaft und Psychologie, Abteilung Berufs- Wirtschafts- und Technikpädagogik.

[15] R. Kosara, C. G. Healey, V. Interrante, D. H. Laidlaw, and C. Ware. *Thoughts on User Studies: Why, How, and When*. IEEE Computer Graphics and Applications, volume 23, 2003.

[16] S. Lee. *Usability Testing for Developing Effective Interactive Multimedia Software: Concepts, Dimensions, and Procedures*. Educational Technology & Society 2(2), 1999.

[17] J. Nielsen. *Usability Inspection Methods*. CHI 94 - Celebrating Interdependence - Conference Companion, 1994.

[18] C. North. *Toward Measuring Visualization Insight*. IEEE Computer Society, 2006.

[19] M. Pohl, M. Schmitt, and S. Diehl. *Comparing the Readability of Graph Layouts using Eyetracking and Task-oriented Analysis*. Computational Aesthetics in Graphics, Visualization, and Imaging, The Eurographics Association, 2009.

[20] L. Prechelt. *Kontrollierte Experimente in der Softwaretechnik*. Springer, Berlin [u.a.], 2001.

[21] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[22] S. Salinger. *Empirische Forschungsmethoden in der Softwaretechnik*. Freie Universität Berlin, 2004.

[23] B. Santos. *Evaluating Visualization techniques and tools: what are the main issues?* Beliv'08, BEyond times and errors: novel evaLuation methods for Information Visualization, Workshop of the ACM Computer Human Conference CHI2008, Florence, 2008.

[24] M. Schiessl, S. Duda, A. Thölke, and R. Fischer. *Eye tracking and its application in usability and media research*. In: MMI-interaktiv Journal - Online Zeitschrift zu Fragen der Mensch-Maschine-Interaktion. Sonderheft: Blickbewegung. 12.03.03, Ausgabe Nr. 6., 2003.

[25] C. Schmoor. *Problematik von Subgruppenanalysen in klinischen Studien - in Methodik klinischer Studien: Methodische Grundlagen der Planung, Durchführung und Auswertung*. Springer, Berlin, 2008.

[26] S. Schnipke and M. Todd. *Trials and Tribulations of Using an Eye-tracking System*. CHI, 2000.

[27] J. Thomas and K. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005.

[28] M. Tory and T. Möller. *Evaluating Visualizations: Do Expert Reviews Work?* IEEE Computer Graphics and Applications, vol. 25, no. 5, S. 8-11, 2005.

[29] J. Tukey. *Exploratory Data Analysis*. Reading, Addison-Wesley, 1977.

[30] A. Uphoff. *Entwicklung einer Abstraktionsschicht zwischen Darstellungen, Eingabeformen und Datenquellen zur visuellen Analyse*. Universität Oldenburg, Fakultät Informatik, Wirtschafts- und Rechtswissenschaften, 2010.

[31] C. Ware. *Visual Thinking: For Design*. Morgan Kaufmann, 2008.

[32] A. Yarbus. *Eye movements and vision*. Plenum Press, New York, 1967.

[33] D. Yoon and N. Narayanan. *Mental Imagery in Problem Solving: An Eye Tracking Study*. ACM, 2004.

Erklärung

Hiermit versichere ich, diese Arbeit selbständig verfasst und nur die angegebenen Quellen benutzt zu haben.

(Natalia Konevtsova)