Navigationsanalyse.

Methodologie der Analyse von Prozessen der Online-Navigation mittels Optimal-Matching

Dissertation
zur Erlangung des Akademischen Grades des
Doktors der Philosophie
am Fachbereich Bildungswissenschaften
der Universität Duisburg – Essen

vorgelegt von

Stefan Iske

geboren in Korbach

Gutachter:
Prof. Dr. Norbert Meder
Prof. Dr. Detlev Leutner

Tag der mündlichen Prüfung: 06.02.2007

Inhaltsverzeichnis

„Der Mensch führt sein Leben und errichtet seine Institutionen auf dem festen Lande. Die Bewegung seines Daseins im Ganzen jedoch sucht er bevorzugt unter der Metaphorik der gewagten Seefahrt zu begreifen. Das Repertoire dieser nautischen Daseinsmetaphorik ist reichhaltig. Es gibt Küsten und Inseln, Hafen und hohes Meer, Riffe und Stürme, Untiefen und Windstillen, Segel und Steuerruder, Steuermänner und Ankergründe, Kompaß und astronomische Navigation, Leuchttürme und Lotsen [...].
Unter den elementaren Realitäten, mit denen es der Mensch zu tun hat, ist ihm die des Meeres – zumindest bis zur späten Eroberung der Luft – die am wenigsten geheuere."

Hans Blumenberg (1997: 9)

„'If in the early years of cinema we already had seminal works that defined the language of the new medium, why haven't we seen the computer-game equivalent of D. W. Griffith *Birth of a Nation*?' The answer, of course, is that we have. The question is how to recognize it."

Mark Tribe. In: Lev Manovich (2001: xiii)

1 Einleitung

Den Ausgangspunkt der vorliegenden Arbeit bildet die Diskrepanz zwischen theoretischen Hypothesen über Navigationsprozesse in Hypertexten und deren empirischer Erforschung.

Seit den Arbeiten der frühen Hypertext-Pioniere Bush (1945), Engelbart (1963) und Nelson (1965) werden das Potenzial und der Mehrwert der Hypertext-Konzeption besonders für den Bereich des Lernens diskutiert. Das Potenzial hypertextueller Umgebungen wird in dessen nicht-linearer Grundstruktur verortet, und aus dieser Grundstruktur wird ein besonderes Potenzial für die Nutzung abgeleitet. Dies öffnet den Raum für eine Vielfalt von Hypothesen über die Aktivitäten von Nutzenden in solchen nicht-linearen Strukturen, die zusammenfassend mit der Metapher der „Navigation“ umschrieben werden.

Analysiert werden diese Navigationsprozesse in der Regel aus *retrospektiver* Perspektive. Ausgangspunkt dabei ist der bereits abgeschlossene Navigationsprozess, der z.B. durch Leistungs- und Vergleichstests zum Gegenstand der Analyse wird. Anhand der *Resultate* von Navigationsprozessen wird dabei auf die *Prozesse* als solche rückgeschlossen. Die Analyse der Navigation in Online-Umgebungen aus *Prozess*perspektive stellt in der gegenwärtigen Forschung zu E-Learning einen blinden Fleck dar.

Auf welcher methodologischen Grundlage können Navigations*prozesse* in Online-Lernumgebungen analysiert werden? Wie können empirisch vorliegende Navigationsverläufe als Sequenzen analysiert werden? Wie können Muster, Regelmäßigkeiten und Strukturen in Navigationssequenzen identifiziert werden? Wie können Navigationssequenzen miteinander verglichen werden? Wie können ähnliche Navigationssequenzen gruppiert werden?

Diese Arbeit gliedert sich in zwei Teile: Im ersten Teil werden die methodologischen Grundlagen der explorativ-heuristischen Analyse von Navigations*prozessen* entwickelt (Kap. 2 – Kap. 8). Im zweiten Teil werden das analytische Potenzial und das konkrete Vorgehen der entwickelten Methodologie anhand der Analyse von Logdaten als Verhaltensspuren demonstriert sowie Ergebnisse dieser Analyse präsentiert (Kap. 9 – Kap. 12). Die empirische Grundlage dieser Studie bilden Navigationssequenzen der Nutzung einer hypertextuellen, metadatenbasierten Online-Lernumgebung (vgl. Kap. 10.1, *Datenerhebung und Datenbasis*).

In *Kapitel 2* werden die *methodologischen Grundlagen* der Navigationsanalyse erläutert. Dabei wird insbesondere das Potenzial einer Triangulation von Ton-, Bild- und Textdaten für die Analyse von Navigations*prozessen* dargestellt. Die in dieser Arbeit entwickelte Navigationsanalyse wird innerhalb der Triangulation

im Bereich der Analyse von Textdaten verortet. Dabei wird der Navigationsprozess als Abfolge der von den Nutzerinnen und Nutzern ausgewählten Seiten einer Lernumgebung analysiert.

In *Kapitel 3* wird anhand der *Web-Didaktik* die Konzeption der Online-Lernumgebung näher beschrieben, die den Ausgangspunkt der durchgeführten Studie darstellt. Insbesondere wird in diesem Kapitel der Navigationsprozess als *autodidaktisches Handeln* gekennzeichnet.

In *Kapitel 4* wird am Beispiel der *Analyse von Logfiles* und des *Web-Mining* der Forschungskontext der Navigationsanalyse dargestellt. Zur Abgrenzung wird die Logdaten-Analyse als Analyse *aggregierter* Logfile-Daten näher ausgeführt und der Analyse *sequenzierter* Logdaten gegenüber gestellt. Anhand des *Data-Mining* wird ein Forschungsfeld skizziert, dass wie die Navigationsanalyse die Handlungen von Nutzenden in Online-Umgebungen analysiert, sich jedoch hinsichtlich des Schwerpunktes, der methodologischen Grundlagen und der Zielsetzung unterscheidet.

In *Kapitel 5* wird die Navigationsanalyse als Analyse von Sequenzen beschrieben. Diese Sequenzen basieren auf den Navigationsverläufen einzelner Nutzer und Nutzerinnen in einer metadatenbasierten, hypertextuellen Lernumgebung. Einleitend wird der Begriff der „Sequenz“ erläutert, daran anschließend wird die Verwendung der Methode der Sequenzanalyse in der soziologischen Lebenslaufforschung skizziert. Dabei wird das analytische Potenzial der Sequenzanalyse für die Analyse von Navigationsverläufen herausgearbeitet.

In *Kapitel 6* wird das *Optimal-Matching Verfahren* als zentrale Methode der Sequenzanalyse erläutert. Anhand der grundlegenden Operationen *Einfügen*, *Löschen* und *Ersetzen / Austauschen* wird die Levenshtein-Distanz bestimmt, die als Maßzahl die Distanz zwischen je zwei Sequenzen ausdrückt. Einleitend wird anhand der Hamming- und der Levenshtein-Distanz auf unterschiedliche Verfahren der Bestimmung der Ähnlichkeit von Sequenzen hingewiesen. Anschließend wird anhand der Gewichtung der grundlegenden Operationen durch *Kosten* eine Spezifizierung der *Optimal-Matching* Analyse vorgestellt. Abschließend wird das Potenzial der *Optimal-Matching* Analyse für die Analyse von Navigationssequenzen zusammengefasst.

In *Kapitel 7* wird zur Verdeutlichung des methodischen Vorgehens der Navigationsanalyse die Funktionsweise der *Optimal-Matching* Analyse anhand eines beispielhaften Datensatzes demonstriert. Dabei werden insbesondere die Effekte unterschiedlicher Kostendefinitionen miteinander verglichen. Die Ergebnisse dieser Optimal-Matching Analyse bilden den Ausgangspunkt einer Clusteranalyse, durch die die Sequenzen hinsichtlich des Kriteriums der Levenshtein-Distanz fusioniert werden. Abschließend werden die Ergebnisse zusammenfassend interpretiert.

In *Kapitel* 8 wird die Sequenzdatenanalyse in ihrem forschungsmethodologischen Kontext verortet, um eine erweiterte Perspektive auf die Methode des Optimal-Matching zu ermöglichen. Mit der *Ereignisdatenanalyse* wird eine weitere Methode der Analyse von Verlaufsdaten dargestellt. Die Ereignisdatenanalyse wird vor allem in der soziologischen Lebenslaufforschung verwendet und zielt auf die Berechnung der *Wahrscheinlichkeit von Zustandswechseln (Übergangswahrscheinlichkeiten)*.

In *Kapitel 9* werden die grundlegenden Entscheidungen des *methodischen Vorgehens* der Navigationsanalyse mittels Optimal-Matching ausgeführt: Als zentrale Elemente der Navigationsanalyse werden die verwendete *Definition der Substitutionskosten,* das verwendete *Clusterverfah*ren sowie die verwendete *Clusteranzahl* formal und inhaltlich auf empirischer Basis begründet.

In *Kapitel 10* werden die *konzeptionell-theoretische* sowie die *programmtechnische Durchführung* der Navigationsanalyse zusammengefasst. Dabei wird insbesondere die Validierung der technischen Erhebung der Sequenzdaten erläutert sowie die Syntax zur Durchführung der Optimal-Matching Analyse mit dem Programm „Transition Data Analysis“ (TDA). Darüber hinaus wird die Datenerhebung in Form von Verhaltensspuren beschrieben, sowie die der durchgeführten Studie zu Grunde liegende empirische Datenbasis.

In *Kapitel 11* werden die *formalen* und *inhaltlichen* Ergebnisse der durchgeführten Navigationsanalyse dargestellt: Welche typischen Abfolgen von Elementen sind in den Navigationssequenzen enthalten? Welche Aussagen können über die empirischen Navigationsverläufe getroffen werden? Welche Muster, Regelmäßigkeiten und Strukturen kommen in den Navigationsverläufen zum Ausdruck? Welche Strategien der Navigation sind identifizierbar? Abschließend werden die Ergebnisse der Interpretation der Navigationssequenzen zusammengefasst und es wird auf *Weiterentwicklungen* und *Variationsmöglichkeiten* der Navigationsanalyse hingewiesen.

In Kapitel 13 werden die zentralen Ergebnisse dieser Arbeit zusammengefasst.

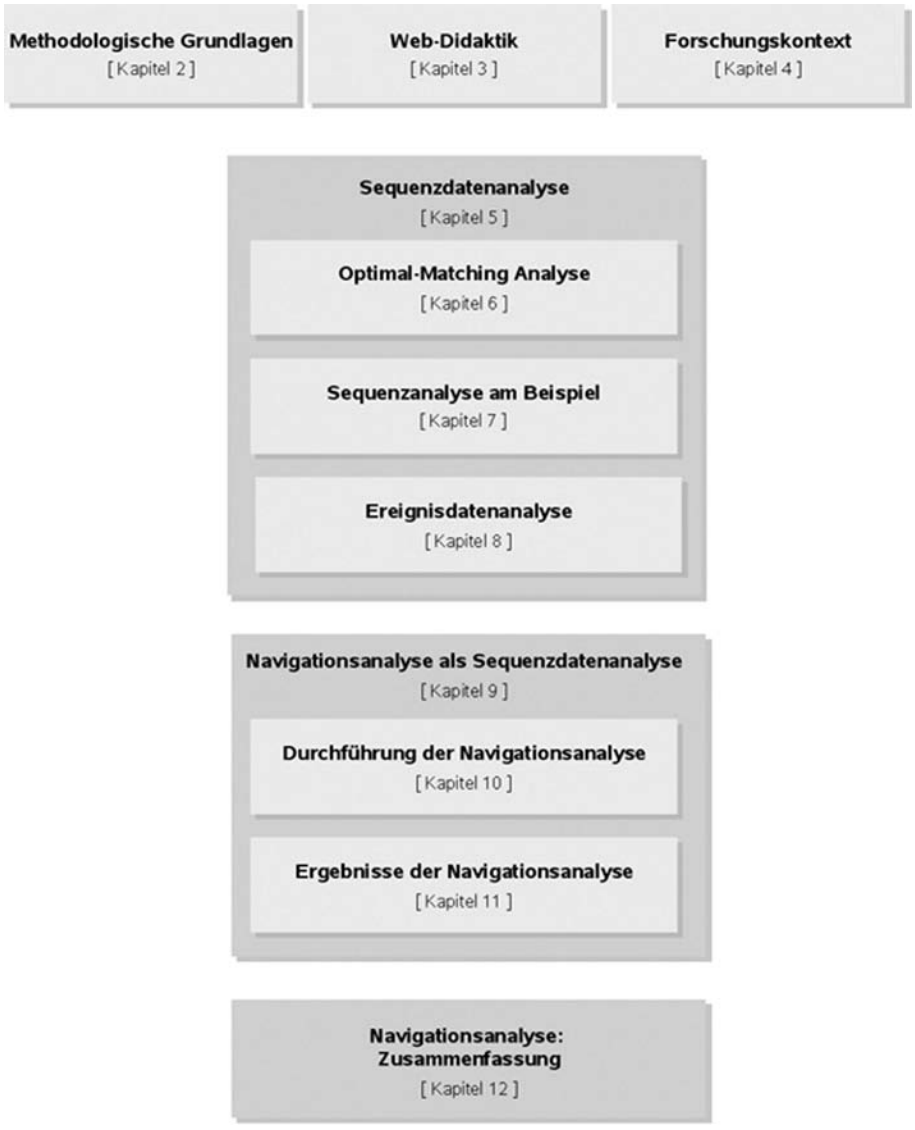

Abbildung 1: Navigationsanalyse im Überblick

2 Methodologische Grundlagen der Navigationsanalyse

In diesem Kapitel werden die methodologischen Grundlagen der Analyse von Navigationsprozessen in Online-Lernumgebungen dargestellt. Den Ausgangspunkt bildet ein triangulativer Ansatz auf der Grundlage von Ton-, Bild- und Textdaten.[1] Dieser triangulative Ansatz bildet den Rahmen der in der vorliegenden Arbeit durchgeführten Navigationsanalyse mittels Optimal-Matching. Mit anderen Worten: die vorliegende Arbeit stellt den Teilbereich einer umfassenden Analyse von Navigationsprozessen dar, der die Analyse von Textdaten in Form sequenzierter Logdaten fokussiert.

Lernende eignen sich bei der Navigation in Online-Lernumgebungen Wissen an. Studien, die diese Aneignung von Wissen untersuchen, stellen dabei häufig das Resultat des Aneignungsprozesses in den Mittelpunkt. Dieser Zugang mit der Frage nach Effektivität und Effizienz dominiert die gegenwärtige Perspektive auf E-Learning – gerade auch im Vergleich zu traditionellen analogen Medien wie dem Buch.

Die Navigationsprozesse von Lernenden *während* des E-Learning bleiben dabei unberücksichtigt. Dass beim E-Learning Aneignungsprozesse stattfinden wird zwar mit Begriffen wie dem *selbstgesteuerten Lernen* vorausgesetzt, gleichzeitig jedoch nur in soweit berücksichtigt, wie sich der Prozess im Resultat wieder findet. Dieser Zugang spiegelt sich in der Methodologie der Untersuchungen wieder: es wird eine *retrospektive Perspektive* auf E-Learning eingenommen. Ansatzpunkt ist der bereits abgeschlossene Aneignungsprozess, der z.B. durch Leistungs- und Vergleichstests zum Gegenstand der Analyse wird.

Aus erziehungswissenschaftlicher Perspektive ist dieser retrospektive Zugang allein unbefriedigend, da Kenntnis von Aneignungsprozessen der Lernenden Aufklärung über pädagogisches Handeln erwarten lässt und darüber hinaus eine Optimierung der Lernumgebung ermöglicht. Für die Analyse von Navigationsprozessen ist daher eine Methodologie erforderlich, mit der neben der äußerlichen Beschreibung auch die impliziten und expliziten Regeln, denen die Lernenden beim Navigieren in einer hypertextuellen Lernumgebung folgen, dargestellt werden können. Die Methodologie muss die *periaktionale Perspektive*, die den Lern*prozess* zum Gegenstand macht, berücksichtigen.

Eine periaktionale Perspektive kann in der Erziehungswissenschaft mit unterschiedlichen Forschungsmethoden eingenommen werden, die jeweils einen anderen Blick auf den Gegenstand kennzeichnen. Ausgangspunkt der Navigationsanalyse ist die empirische Perspektive mit den grundlegenden Fragen, welchen Einfluss die Aufzeichnungsverfahren auf die empirische Untersuchung der Aneignung von Wissen in Online-

1 Die methodologischen Grundlagen der Analyse von Navigationsprozessen wurden in Zusammenarbeit mit Christian Swertz entwickelt (vgl. Iske / Swertz 2005).

Lernumgebungen haben und mit welchen Verfahren die Aneignungsprozesse angemessen untersucht werden können.

In didaktischer Perspektive bildet sich im multisequenziellen Nutzungspfad der Online-Lernenden eine *Autodidaktik* (vgl. Meder 2003: 52) ab, bei der Geltung in den zeitlichen Verlauf der Navigation abgebildet wird (vgl. Kap. 3.2). Nun kann bezweifelt werden, dass die subjektiven Zwecke der autodidaktisch Handelnden überhaupt mit empirischen Methoden abbildbar sind. Inwiefern ist eine solche Abbildung möglich?

Aneignungsprozesse zu untersuchen setzt stets voraus, dass auch die Forscherinnen und Forscher Aneignungsprozesse durchlaufen sind. Schon daran wird erkennbar, dass Bildungswissenschaft es nicht mit kausalen Beziehungen zu tun haben kann, sondern der Gegenstand als sinnhafte Beziehung zu kennzeichnen ist (vgl. Hönigswald 1927: 48). Ziel der empirischen Bildungsforschung können daher nicht universelle Gesetze oder Regeln sein. Vielmehr geht es um eine Rekonstruktion sinnhaften Verhaltens (vgl. Bohnsack 1999), bei der das sinnhafte Verhalten der Forscherinnen und Forscher berücksichtigt werden muss.

Der Ausdruck von Aneignungsprozessen erfolgt nun stets in Medien. Medien sind Gegenstände, die von Menschen zu Zeichen gemacht werden, z.B. Gesten, Sprache, Schrift usw. (vgl. Swertz 2001). Medien weisen nach diesem Verständnis stets eine physikalische, eine semiotische und eine pragmatische Dimension auf. Eine Eigenschaft von Medien ist die mediale Reflexivität (vgl. McLuhan 1995). Eine mediale Reflexion ist die Spiegelung eines Raumes in einem Aspekt. In einem Medium reflektieren die physikalische, die semiotische und die pragmatische Dimension aufeinander. Ein Beispiel: In der Musik ist der Laut nur ein Aspekt aus dem Möglichkeitsraum der Luft. Dennoch spiegeln sich in dem Laut die physikalischen Eigenschaften der Luft. In der semiotischen Dimension ist ein Wort nur eine Auswahl aus dem Möglichkeitsraum der Sätze. Dennoch spiegeln sich in dem Wort die semiotischen Eigenschaften der Sätze. In der pragmatischen Dimension ist die mediale Äußerung nur ein Aspekt aus dem Raum möglicher Äußerungen eines Subjekts. Dennoch spiegeln sich in der medialen Äußerung die Eigenschaften des Subjekts. Mediale Äußerungen sind nicht identisch mit Subjekten, aber diese medialen Äußerungen verweisen auf das Subjekt.

Aneignungsprozesse können also rekonstruiert werden, wenn sich das Subjekt in medialen Äußerungen ausdrückt und die Forscherinnen und Forscher selbst Aneignungsprozesse durchlaufen sind. Damit ist die Möglichkeit der Rekonstruktion von Aneignungsprozessen anhand empirischer Daten gezeigt und zugleich die Reichweite der Analyse begrenzt.

Neben dieser systematischen Grenze besteht für die empirische Forschung eine weitere Grenze in den Verfahren der Datenerhebung: Gegenstand der Forschung wird nur, was dauerhaft fixiert werden kann. Insofern setzt eine empirische Untersuchung von Aneignungsprozessen in Medien immer schon den Gebrauch von (Aufzeichnungs-)Medien voraus. Mit den Aufzeichnungsverfahren wird festgelegt, welche Äußerungen eines Subjekts als Grundlage der Rekonstruktion von Aneignungsprozessen verwendet werden können – und welche durch die Eigenschaften der Aufzeichnungsgeräte ausgeschlossen werden. Den Ausgangspunkt der

Rekonstruktion von Navigationsprozessen bildet auf methodologischer Ebene eine Datentriangulation auf der Grundlage von Text-, Ton- und Bilddaten.

Um diese Methodologie der Navigationsanalyse diskutieren zu können ist angesichts der Vielfalt von Online-Lernumgebungen zunächst zu klären, in welcher Art von Online-Lernumgebung die Aufzeichnung erfolgen sollte. Auf technologischer Ebene sind Online-Lernumgebungen Hypertexte. Auf zeitlicher Ebene handelt es sich bei Aneignungsprozessen in hypertextuellen Online-Lernumgebungen um das lineare Entfalten eines nicht-linearen Hypertextes (Kuhlen 1991: 33) in den Verlauf der Bearbeitungszeit. So bezeichnet Landow (1997) Hypertexte als *multilinear* bzw. *multisequenziell*: Durch die Navigation, d.h. die Auswahl bestimmter Verknüpfungen aus einer Vielzahl möglicher Verknüpfungen, entsteht ein sequenzieller Nutzungspfad. Häufig reduziert sich die Funktionalität von E-Learning-Plattformen jedoch auf die Bereitstellung von Dokumenten zum Download. Diese Form des Online-Lernens wird von Jeschke / Keil-Slawik (2004) zu Recht als reines „Dokument Management Center“ kritisiert, da der eigentliche Lernprozess aus der Lernumgebung ausgeklammert und in die bereitgestellten Dokumente verlagert wird. Aufzeichnungen des Navigationsverhaltens in einer solchen Distributionsplattform sagen daher wenig über Aneignungsprozesse aus. Daher ist es erforderlich, eine Online-Lernumgebung zu verwenden, die den Lernprozess in der Plattform ermöglicht. Das ist insbesondere der Fall, wenn das Lernmaterial nicht als für den Druck formulierter Fließtext, sondern als modularisierter Hypertext bereitgestellt wird (vgl. Iske 2002). Ausgehend von diesen Überlegungen bildet die Lernumgebung *Lerndorf* (vgl. Swertz 2004) den Ausgangspunkt der vorliegenden Navigationsanalyse. Die didaktische Konzeption dieser Lernumgebung wird in Kapitel 3 dargestellt.

2.1 Aufzeichnungsverfahren

Die folgende Darstellung der Aufzeichnungsverfahren im Rahmen der Navigationsanalyse beschränkt sich auf Ton-, Text- und Bilddaten. Diese Einschränkung ist nicht systematisch begründet, sondern pragmatisch durch die verfügbare Technik bedingt. So ist z.B. die Aufzeichnung haptischer oder olfaktorischer Daten derzeit kaum möglich (auch wenn eine solche Aufzeichnung relevant sein könnte). Für einen grundlegenden Überblick werden alle drei Datentypen dargestellt. Der Fokus der vorliegenden Arbeit liegt jedoch auf der Analyse von Logfile-Daten als spezieller Form von Textdaten.

2.1.1 Tondaten

Tondaten fallen bei der Navigationsanalyse dann an, wenn mittels eines Mikrofons und eines Aufnahmegeräts die Geräusche während des Navigationsprozesses festgehalten werden. Die Aufzeichnungen sind in

einer Qualität möglich, mit der z.B. Störungen von außen, Räuspern oder eine Veränderung im Tonfall abgebildet werden können. Welche Geräusche dabei hervorgehoben werden hängt in hohem Maße von der Position und der Art des Mikrofons ab. So ist z.B. die Aufzeichnung der Geräusche in einer Selbstlernphase in einem mit 40 Lernenden besetzten Computerraum möglich, die einzelnen Gespräche bleiben jedoch unverständlich. Die Verwendung eines Mikrofons fokussiert damit die Perspektive auf den einzelnen Lernenden.

Bei der Beobachtung einzelner Lernender können die unbewussten Laute für die Interpretation von Aneignungsprozessen verwendet werden, da implizite Regeln in Erstaunen, Zustimmung, Ablehnung etc. zum Ausdruck kommen. Da Aneignungsprozesse jedoch in hohem Maße an Sprache gebunden sind, bleibt eine solche Analyse unbefriedigend, insbesondere da Äußerungen über explizite Regeln der Aneignungsprozesse auch in Sprache ausgedrückt werden können.

Eine Möglichkeit zur periaktionalen Erhebung von Verbalisierungen ist die Methode des *Lauten Denkens* (vgl. „Protocol Analysis“, Ericsson / Simon 1999, 1980). Das besondere Potential des *Lauten Denkens* für die Untersuchung von Navigationsprozessen liegt dabei im zeitlichen Zusammenfallen des *Handelns* und des *Verbalisierens*, d.h. im *gleichzeitigen* Verbalisieren *während* der Bearbeitung einer Aufgabe. Diese Verbalisierungen stehen somit im Zusammenhang mit beobachtbarem Handeln. In methodischer Perspektive ist darauf hinzuweisen, dass bei diesen verbalen Berichten vor allem Informationen über den Prozess bzw. die Sequenz der Aufgabenbearbeitung im Mittelpunkt stehen. Der Aspekt der zeitlichen Ordnung des Bearbeitungsprozesses tritt dabei in den Vordergrund. Dieses Verfahren kann besonders in der nicht-routinisierten, ersten Begegnung mit einer Lernumgebung als angemessenes Verfahren der Datenerhebung angesehen werden (vgl. Oostendorp u. a. 1999). Durch den Einsatz der Methode des *Lauten Denkens* werden vor allem Rationalisierungen und Interpretationen der Lernenden *nach Beendigung* des Aneignungsprozesses vermieden. Das *Laute Denken* kann darüber hinaus mit retrospektiven Befragungsmethoden wie z.B. dem Leitfadeninterview verbunden werden.

Grundlage der Analyse von Tondaten im Rahmen der Navigationsanalyse ist eine Weiterentwicklung des von Ericsson / Simon (1980, 1999) vorgeschlagenen psychologischen Modells der „Protocol Analysis“ für den Bereich der Pädagogik. Methodologische Veränderungen und Erweiterung basieren dabei vor allem auf der Berücksichtigung kommunikations- und sprachwissenschaftlicher Theorien (Yngve 1970; Sacks / Schegloff / Jefferson 1974; Duncan / Fiske 1977; Schegloff 1983; Drummond / Hopper 1993; Jefferson 1993). Darüber hinaus werden Erweiterungen und Veränderungen des Rahmenmodells notwendig, die sich aus der Verbindung der Methode des *Lauten Denkens* mit Computertechnologie ergeben, d.h. aus der Bearbeitung von Aufgaben vor einem Computerbildschirm (vgl. Boren / Ramey 2000).

Ausgehend von diesen Überlegungen können durch die Verwendung von Tondaten subjektive Zwecke erfasst werden. Die Rekonstruktion stößt jedoch dort an Grenzen, wo sich Äußerungen beispielsweise auf eine Bildschirmdarstellung beziehen, die ihm Rahmen der Tonaufzeichnung nicht dokumentiert wird. In solchen Fällen sind Bild- und Textdaten zur Kontextualisierung der Tondaten erforderlich.

2.1.2 Textdaten

Textdaten, in denen der Nutzungspfad der Lernenden festgehalten wird, fallen bei Online-Lernumgebungen in Form von Logdaten an. Logdaten stellen eine spezifische Art der Aufzeichnung und Transkription des Navigationsprozesses dar. Durch die Logdaten wird eine Abfolge von Seiten (der Navigationsverlauf) in eine digitale Textdatei übersetzt, also von im Voraus festgelegten Algorithmen transkribiert. Es handelt sich dabei um eine formale, automatisierte Beschreibung bzw. Analyse visueller Daten (Sequenz der Webseiten, d.h. der Bilder auf dem Bildschirm). Diese formalen Regeln entsprechen genau den durch die Logdaten protokollierten Reaktionen des Webservers. In medialer Hinsicht sind Logdaten Textdaten, die von einem Internet-Server während des Navigationsprozesses automatisch generiert und aufgezeichnet werden.[2] Bergmann / Meier (2000: 431) bezeichnen diese Textdaten daher als „elektronische Prozessdaten". Es handelt sich dabei um eine passive Protokollierung. Priemer (2004) nennt als Vorteile dieser Protokollierung u.a. die unbemerkte und detailgenaue objektive Aufzeichnung ohne Beeinflussung der Nutzenden in authentischen Nutzungssituationen.

Die Logdaten werden nicht nur automatisch aufgezeichnet, sondern können auch automatisch ausgewertet werden. Die automatische Aufzeichnung und Auswertung fokussiert die Perspektive auf große Lerngruppen. Aufgezeichnet, d.h. gemessen werden dabei technische Transaktionsdaten. Eine Analyse dieser technischen Transaktionsdaten bleibt, wie Huber / Mandl (1994: 12) zutreffend bemerkt haben, oberflächlich. Allerdings enthalten die Transaktionsdaten eindeutige Referenzen auf die abgerufenen Inhalte. Durch Berücksichtigung der abgerufenen Inhalte können die impliziten Regeln der Autodidaktik in Bezug auf die sachlogische Navigation rekonstruiert werden (vgl. Kap. 3.2, *Navigation als Autodidaktik*).

Zwar ist dem Einwand von Wirth / Brecht (1999: 153) zuzustimmen, dass die Interpretation der in den Logdaten enthaltenen Navigationsmuster erschwert wird durch die fehlende Einsicht in die Intentionen der Lernenden, die den Navigationshandlungen zu Grunde liegen. Bezieht man die aufgerufenen Inhalte ein (z.B. in Form von entsprechenden Metadaten), können durchaus Intentionen dargestellt werden. Diese bleiben jedoch gegenüber den Intentionen der Lernenden notwendig distanziert. Diese Distanz wird im Rahmen der Navigationsanalyse durch die Datentriangulation verringert. Die Analyse von Logdaten in Form sequenzierter Daten (Sequenzen) wird ausführlich in Kapitel 5, *Sequenzdatenanalyse* dargestellt.

2.1.3 Bilddaten

Verfahren der audiovisuellen Verhaltensdokumentation bezeichnet Bergmann (1985) als „audiovisuelle Reproduktionsmedien". Die technische Fixierung ermöglicht eine *registrierende Konservierung*, die technisch

2 Das Logfile enthält beispielsweise Informationen darüber, welche Objekte angefordert werden, welcher Computer (IP) die Objekte anfordert oder ob die Übertragung der Daten erfolgreich war.

automatisiert und ohne Interpretation abläuft. Bergmann distanziert sich von der Auffassung der *registrierenden Konservierung* als einer reinen Abbildung der Wirklichkeit und weist auf dessen konstruktiven Charakter und dessen hergestellte Authentizität hin.[3] Ein wesentliches Motiv für die Verwendung von registrierenden Konservierungstechniken in der Forschung sieht Bergmann darin, dass sie es erlauben, den Ablauf und die Sinnstruktur eines soziales Geschehen in den Blick zu nehmen und zu analysieren.

Für eine Navigationsanalyse unter der Perspektive von Aneignungsprozessen ist der Einsatz audiovisueller Reproduktionsmedien von großer Bedeutung, da die Interaktionen der Lernenden mit der Lernumgebung als prozessuales Geschehen den Gegenstand bilden. Zur Analyse des Aneignungsprozesses bietet die Verwendung von audiovisuellen Reproduktionsmedien daher ein besonderes Potenzial: Der Prozess wird in seinem Ablauf registriert und vermittels der Registrierung zum Gegenstand der Beobachtung und Analyse.

Damit ist eine Verschiebung von Messung und Interpretation verbunden. Während bei der teilnehmenden Beobachtung oder einem Interview Messvorgang und Interpretation nicht voneinander zu trennen sind, kann durch die Aufzeichnung die Messung und die Interpretation voneinander getrennt werden. Bergmann (1985) spricht in diesem Zusammenhang von einem „primären Sinnzusammenhang“ als sich vollziehendes Geschehen und einem „sekundären Sinnzusammenhang“ als Darstellung bzw. Interpretation eines vollzogenen Geschehens. Dabei ist der Messvorgang nicht interpretationsfrei. Durch die Verlagerung in die Technik ist nach der Festlegung der Perspektive und des Verfahrens die Messung jedoch mit einem unabhängig von den Forschenden sich stets identisch wiederholenden Ablauf verbunden. Mit diesem Vorteil ist gleichzeitig der Nachteil verknüpft, dass die Interpretation nur in der durch die Messung eingeschränkten Perspektive erfolgen kann. Die Wahl der Perspektive muss daher bei der Interpretation berücksichtigt werden, indem z.B. expliziert wird, welche Prozesse nicht berücksichtigt werden konnten, weil sie durch die Perspektiventscheidung ausgeschlossen wurden.

Als methodischen Vorteil der Verwendung von (elektromagnetischen) Aufzeichnungen nennt auch Bohnsack (1999: 144) die Trennung von Datenerhebung und Interpretation, da bei der Datenerhebung die Daten nicht durch die Sprache eines Beobachters gefiltert werden und durch die Aufzeichnung der nachträglichen Transkription und Interpretation zugänglich sind. Auf forschungsmethodologischer Ebene ermöglicht dies vor allem intersubjektive Überprüfbarkeit der Interpretationsleistung des Beobachters. Dabei übersieht Bohnsack jedoch, dass auch die Aufzeichnung eine Interpretation darstellt, die z.B. durch die Sprache eines Beobachters gefiltert wird. Im Fall der Navigationsanalyse drückt der Beobachter seine Beobachtungsabsicht im verwendeten Algorithmus aus, der die Grundlage der Erstellung der Logdaten bildet.

3 „Denn was heißt hier ‚wirklich‘? Zur ‚Wirklichkeit‘ eines sozialen Geschehens gehört ja gerade das, was seine methodologische Fixierung für die soziologische Analyse notwendig macht - seine Flüchtigkeit. Aber eben mit seiner Fixierung büßt ein soziales Geschehen seine Flüchtigkeit ein. Demnach ist die audiovisuelle Aufzeichnung eines sozialen Geschehens keineswegs die rein deskriptive Abbildung, als welche sie zunächst erscheinen mag, ihr ist vielmehr in ihrer zeitmanipulativen Struktur grundsätzlich ein konstruktives Moment eigen“ (Bergmann 1985: 317).

2.2 Triangulation von Text-, Ton- und Bilddaten

Es wurde deutlich, dass die Aufzeichnung von Text-, Ton- und Bilddaten eine geeignete Grundlage für die Analyse von Navigationsprozessen darstellt. Es ist davon auszugehen, dass die Interpretation in Verbindung dieser drei Datentypen eine genauere Annäherung an Aneignungsprozesse ermöglicht als die Interpretation der jeweils einzelnen Daten. Diese Rekonstruktion mit Hilfe unterschiedlicher Aufzeichnungsverfahren dient dabei weniger als Strategie der Validierung der Ergebnisse der einzelnen Methoden[4], sondern zielt darauf ab, die Analyse auf eine breitere Basis zu stellen.

Ausgangspunkt der Triangulation ist hier der Umstand, dass die erhobenen Bild-, Ton- und Textdaten eine je spezifische mediale Perspektive auf ein und denselben Aneignungsprozess liefern und durch dieses komplementäre Verhältnis eine dichte Darstellung ermöglichen. Die Relationierung der Perspektiven ist damit ein wichtiger Schritt der Analyse:

- Die Textdaten (Logdaten) werden im Kontext der Tondaten (Verbalisierung) und der Bilddaten (Handlungen) analysiert und interpretiert;
- Die Tondaten (Verbalisierung) werden auf im Kontext der Textdaten (Logdaten) und Bilddaten (Handlungen) analysiert und interpretiert;
- Die Bilddaten (Handlungen) werden im Kontext der Tondaten (Verbalisierung) und der Textdaten (Logdaten) analysiert und interpretiert.

Betont wird damit der komplementäre und divergente[5] Charakter der durch die unterschiedlichen Methoden erhaltenen Daten.[6] Aufschlussreich sind demnach bei der Datentriangulation sowohl Übereinstimmungen als auch Differenzen[7], wobei die letzteren in besonderer Weise nach zusätzlichen theoretischen und empirischen Klärungen verlangen (vgl. Flick 2004, Marotzki 1999).

Steht die Rekonstruktion und Analyse von Aneignungs- und Verstehensprozessen im Vordergrund, reicht der Rückgriff auf aggregierte statistische Kennzahlen des Navigationsverhaltens wie Zugriffsstatistiken, auf physiologische Messungen oder auf *Eyetracking* nicht aus (vgl. Huber / Mandl 1994). Diese Daten lassen bestenfalls in sehr eingeschränkter Weise Antworten auf die eingangs gestellten Fragen zu. In Anlehnung an Olson et al. (1984: 254) stehen bei der Navigationsanalyse weniger statistische Kennzahlen im Vordergrund, die aus aggregierten Logdaten errechnet werden, sondern vielmehr die Aneignungs- und Verstehensprozesse, die diesen zugrunde liegen. Diese Prozesse bilden den Gegenstand der Navigationsanalyse.

4 Zur Kritik am Konzept der Triangulation als Validierung, vgl. Flick (2004: 18f.); Denzin / Lincoln (1994).

5 „In der ethnographischen Forschungspraxis führt die Triangulation von Datentypen und Methoden sowie von theoretischen Perspektiven zu erweiterten Erkenntnismöglichkeiten, die sich aus Konvergenzen aber mehr noch aus den Divergenzen, die sie hervorbringen bzw. produzieren, speisen“ (Flick 2004: 66).

6 So verweist z.B. Bohnsack (1999: 146) auf die Vorteile der komplementären Verwendung der teilnehmenden Beobachtung und des Interviews.

7 Vgl. „reflexive Triangulation“ (Hammersley / Atkinson 1983).

In der Absicht einer methodologischen Triangulation basiert die Navigationsanalyse auf dem Einsatz unterschiedlicher Methoden der Datenerhebung (vgl. Denzin 1970).

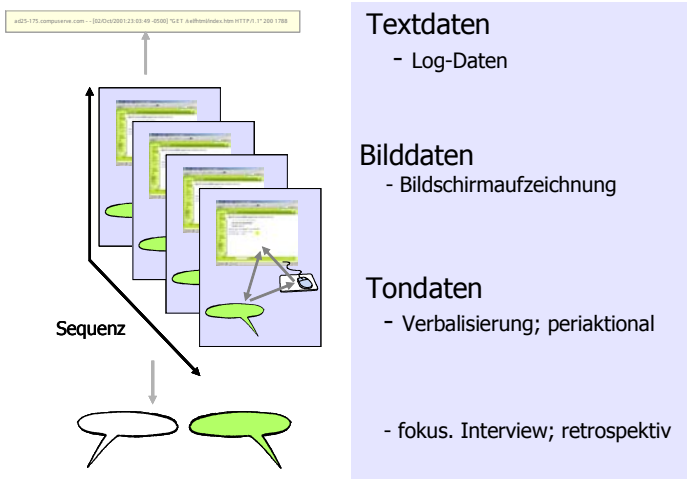

Abbildung 2: Navigationsanalyse: Text-, Bild- und Tondaten

Die im Rahmen der Navigationsanalyse erhobenen Daten bilden den Prozess der Navigation einschließlich einzelner Schritte ab. Sie werden sowohl auf der Ebene der Handlung als auch auf der Ebene der sprachlichen Äußerungen rekonstruierbar und analysierbar. Die Methodologie der Navigationsanalyse als Kombination von Bild-, Ton- und Textdaten fokussiert damit genau den Bereich des E-Learning, der bei der gegenwärtig vorherrschenden Analyse in Form aggregierter Logdaten-Analyse unterbelichtet bleibt: den *Prozess* der Navigation als Prozess der Interaktion mit einer Online-Umgebungen.

3 Web-Didaktik

In diesem Kapitel wird die Konzeption der Lernumgebung[8] näher beschrieben, die den Bezugspunkt der vorliegenden Navigationsanalyse darstellt. Die in dieser Lernumgebung erhobenen Navigationssequenzen bilden die empirische Datenbasis der Navigationsanalyse mittels Optimal-Matching (vgl. Kap. 10.1; *Datenerhebung und Datenbasis*).

Die Berücksichtigung der Konzeption der Lernumgebung ist besonders wichtig, weil ein genereller Zusammenhang zwischen Konzeption / Realisierung von Lernumgebungen und potentiellen sowie empirischen Navigationsverläufen besteht. In Kapitel 2 wurde bereits mit der Kritik an Online-Plattformen als „Dokument Management Center" (Jeschke / Keil-Slawik 2004) darauf hingewiesen, dass eine solche Strukturierung von Plattformen nicht zur Analyse von Aneignungssprozessen geeignet ist.

Die Frage, welche Lernumgebung für eine Navigationsanalyse besonders geeignet ist, wird in diesem Kapitel unter Bezugnahme auf die *Web-Didaktik* von Meder (2006) beantwortet. Dabei kann die Konzeption der Web-Didaktik nicht umfassend dargestellt werden, sondern lediglich die für die hier vorliegende Navigationsanalyse relevanten Bereiche der didaktischen Ontologie, der rezeptiven Wissenseinheiten und der Mikronavigation als autodidaktischem Handeln.

3.1 Grundlagen der Web-Didaktik

Die Web-Didaktik stellt eine Spezifikation Allgemeiner Didaktik unter den Bedingungen der medialen Strukturen des Word Wide Web dar (vgl. Meder 1995, 1995a, 1998, 2006; Swertz 2001, 2003). Die grundlegende mediale Struktur des WWW ist die Hypertext-Konzeption (Berners-Lee / Cailliau 1990; Berners-Lee 2000; vgl. Iske 2001), die nach Kuhlen (1991) darin besteht, einen Gegenstandsbereich in einzelne Informationseinheiten zu gliedern (Dekontextualisierung) und diese Einheiten durch Verknüpfungen zu verbinden (Rekontextualisierung).

Die Web-Didaktik ist eine metadatenbasierte, pädagogisch-didaktische Interpretation dieser Hypertext-Konzeption, wobei die De- und Rekontextualisierung unter explizit pädagogisch-didaktischer Perspektive statt-

8 Bei der Lernumgebung handelt es sich um <www.lerndorf.de> (vgl. Abbildung 3: 14), und dort um den Bereich *Statistik* mit den Lerneinheiten *Maße der zentralen Tendenz*, *arithmetisches Mittel*, *Modus* und *Median*. In dieser Lernumgebung ist die didaktische Ontologie von Meder (2006) in Hinblick auf die rezeptiven Wisseneinheiten implementiert.

finden: sowohl die Informationseinheiten (Wissenseinheiten) als auch deren Verknüpfungen (Relationen) werden durch didaktische Metadaten typisiert.[9]

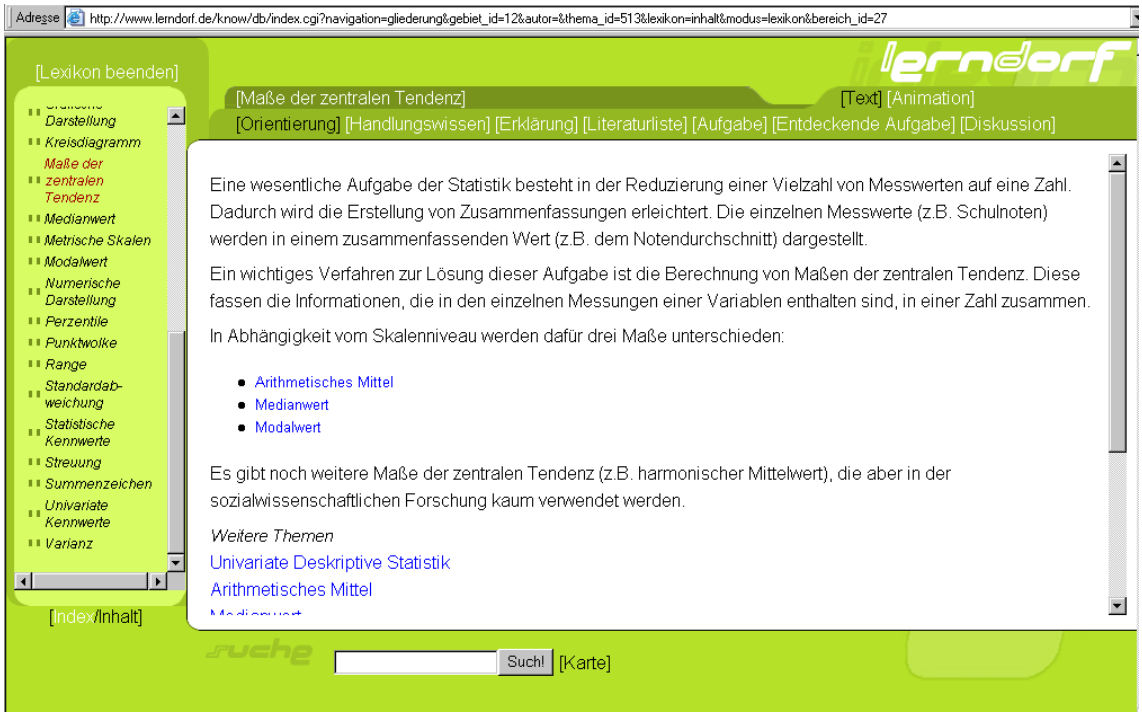

Abbildung 3: Screenshot der Lernumgebung 'Lerndorf' (<www.lerndorf.de>, 29.08.2006)

Den Kern der Web-Didaktik bildet eine didaktische Ontologie[10], d.h. ein systematisch begründetes Klassifikationssystem zur Bestimmung didaktischer Objekte, das sowohl den Autor in der Entwicklung optimaler Lehr-Lern-Wege unterstützt als auch die Lernenden bei der selbstorganisierten Erarbeitung von Wissen: es geht um eine didaktische sowie um eine autodidaktische Unterstützung. Dabei liegt der Web-Didaktik das Bildungsideal eines selbstbestimmten und selbstorganisierten Lernenden zu Grunde, eines „Autodidakten“ (Meder 1987, 2004). Die didaktische Ontologie bildet dabei ein Alleinstellungsmerkmal der Web-Didaktik. Nach Meder (2006: 49f.) verbindet die Web-Didaktik für Online-Lernumgebungen das Potenzial des WWW mit der didaktisch-wissenschaftlichen Tradition der Strukturierung und Gestaltung von Lernmaterialien und Lernprozessen.[11]

9 Zur Typisierung von Verknüpfungen, vgl. Meder (2006: 233f.).

10 „Unter einer *didaktischen Ontologie* versteht man die Klassifikation didaktischer Objekte (das sind im Wesentlichen Lernmaterialien, Aufgaben und Szenarien der Zusammenarbeit) nach didaktischen Kategorien – wie Stoff, Sozialform, medialer Form u.a.m. - sowie die Klassifikation der Relationen, die zwischen den didaktischen Objekten bestehen“ (Meder 2006: 40).

11 Theoretisch schließt die Web-Didaktik an unterschiedliche Traditionen an (vgl. Meder 2006: 39): an die Tradition der Wissensorganisation in der Didaktik (Flechsig 1990; Haller 1995), an den Ansatz des Instructional Design (Merrill 1992, Merrill 1994); an die Tradition der indirekten Erziehung (Rousseau 2001); an die Tradition des Arrangements von Lernumgebungen (Rousseau 2001, Lewin 1982, Winnefeld 1971); an die Tradition der theoretischen Rhethorik (Aristoteles).

Die didaktischen Metadaten werden dabei einerseits verwendet, um bei Bedarf die Lernumgebung an die Lernenden zu adaptieren, andererseits zur Strukturierung des Wissens, um die Lernenden bei der zielgerichteten Navigation durch die Lernumgebung zu unterstützen: die didaktische Ontologie ermöglicht ein „entdeckend-lernendes Surfen in einer Umgebung, die es dem Lernenden über eine professionelle Organisation des Wissens möglich macht, das entdeckende Lernen selbst zu steuern“ (Meder 2006: 23).

Die Grundsätze der Allgemeinen Didaktik fasst Meder in *drei Hauptsätzen der Didaktik* zusammen, die auch die Grundlage der Web-Didaktik darstellen:

- *Erster Hauptsatz*: „Didaktisches Handeln ist die Abbildung von Bedeutungsbeziehungen in die Zeit des lernenden Vollzuges, in die Zeit der Aneignung von Wissen unter dem Gesichtspunkt der Geltung“ (Meder 2006: 35).
- *Zweiter Hauptsatz*: „Die Aneignung von Wissen, das stets aus Bedeutungsbeziehungen besteht, ist die Abbildung der zeitlichen Strukturen des Lernprozesses in den sachlogischen Raum eben dieser Bedeutungsbeziehungen“ (Meder 2006: 36).
- *Dritter Hauptsatz*: „Eine Lernumgebung ist die Abbildung der Bedeutungsbeziehungen in alle bekannten zeitlichen Strukturen des Lernens, so dass jeder Lernende sich den zeitlichen Verlauf, den Weg, wählen kann, der ihn am besten darin unterstützt, seinen Lern- und Bildungsprozess in den logischen Raum der Bedeutungsbeziehungen – in sein mentales Modell – zu transformieren“ (Meder 2006: 37).

Der *erste Hauptsatz* schließt unmittelbar an die Arbeiten von Richard Hönigswald (1913, 1927) an, der in systematisch-theoretischer Perspektive Didaktik als die Abbildung von Bedeutungsbeziehungen in die Zeit definiert: Bei den Bedeutungsbeziehungen handelt es sich um sachlogische Beziehungen, die als solche keine zeitliche Struktur aufweisen. Zum Zweck der Tradierung müssen jedoch diese sachlogischen Beziehungen in eine zeitliche Struktur abgebildet werden, in die Zeit des Erlebens und des Lernens. Verdeutlicht werden kann der *erste Hauptsatz* am Beispiel der *Unterrichtsplanung* von Lehrenden:[12] Diese führen zunächst eine *Sachanalyse* durch, um den *Unterrichtsgegenstand* als Struktur von Bedeutungen und deren Beziehungen zu erschließen. Dieser Unterrichtsgegenstand ist in der Regel *sachlogisch* strukturiert und vieldimensional. In der anschließenden *Verlaufsanalyse* erstellen die Lehrenden auf Grundlage der Sachanalyse eine *Unterrichtsplanung* als Reihenfolge des unterrichtlichen Vorgehens. Diese Unterrichtsplanung ist als konkretes Vorgehen in der Unterrichtssituation immer *zeitlich* strukturiert und zielt auf die angemessenste und geeignetste Reihenfolge und Form der Behandlung der Inhalte, um bei einer spezifischen Zielgruppe Lernen zu ermöglichen.[13] Der Gesichtspunkt der *Geltung* verweist in diesem Fall z.B. auf *Curricula* im Be-

12 „Jeder, der vor der Aufgabe der Unterrichtsplanung steht, kennt das Problem: Nach der Sachanalyse des Unterrichtsstoffes, die eine räumliche sachlogische Struktur ergibt (in der Regel ein semantisches Netz, einen Begriffsbaum oder eine Matrix), muss man zur Verlaufsanalyse übergehen und die räumliche Struktur zeitlich linearisieren.“ (Meder 2006: 33).

13 „Das ist im übrigen kein triviales Geschäft, wenn man bedenkt, dass der Lernstoff in der Regel die Komplexität einer vernetzten Struktur hat und dass im didaktischen Handeln diese vernetzte Struktur in eine lineare Struktur übersetzt werden muss. Denn Lernverläufe sind immer linear; sie verlaufen in der Zeit der Aneignung [...]“ (Meder 2006: 27).

reich der Schule, in denen die zu vermittelnde Inhalte und die zu vermittelnden Kompetenzen dokumentiert werden. Übertragen auf den Bereich des Online-Lernens definiert der *erste Hauptsatz* den Gegenstand der Web-Didaktik als Abbildung von Bedeutungsbeziehungen in die Zeit des E-Learning, d.h. diese Abbildung muss grundsätzlich in einer Online-Lernumgebung möglich sein.

Während sich der *erste Hauptsatz* der Didaktik auf die Perspektive eines Lehrenden als Organisator oder als Organisatorin des Lehrprozesses beziehen lässt, verdeutlicht der *zweite Hauptsatz* komplementär die Perspektive der Lernenden und der konkreten Lernprozesse. Diese eignen sich Wissen an, indem sie den zeitlich verlaufenden Lernprozess in den sachlogischen Raum der Bedeutungsbeziehungen abbilden, also auf Grundlage des zeitlich verlaufenden Lernprozesses ein mentales Modell der Bedeutungsbeziehung entwickeln. Um im obigen Bild zu bleiben: Die Schülerinnen und Schüler eignen sich Wissen an, indem sie ausgehend vom unterrichtlichen Vorgehen als zeitlich-linearem Verlauf ein mentales Modell des Gegenstandsbereichs (der Bedeutungsbeziehung) entwickeln, das nicht zeitlich-linear sondern sachlogisch strukturiert ist.

Der *dritte Hauptsatz* definiert die Anforderungen an eine Lernumgebung. Diese bestehen in der Bereitstellung eines Möglichkeitsraums, in dem die Abbildung der Bedeutungsbeziehungen auf vielfältige Weise möglich ist.[14] So können z.B. in die Lernumgebung didaktische Heuristiken als vordefinierte Wege des Lernens implementiert und zur Verfügung gestellt werden (*guided tour*, *Kurs*). Die Lernenden können dann einerseits den Lernweg als zeitlichen Verlauf auswählen, der sie am besten bei ihren Lernprozessen unterstützt, oder aber die Lernumgebung frei explorieren und damit selbst über ihre Abbildungsprozesse entscheiden. Im dritten Hauptsatz wird die Tradition der *indirekten Erziehung* und des *Arrangements von Lernumgebungen* als Wurzeln der Web-Didaktik deutlich. Im Gegensatz zum Unterricht als einer spezifischen Form von Lernumgebungen für das Lernen in der Gruppe unterstützen Online-Lernumgebungen nach der obigen Anforderung individualisiertes und zielgerichtetes Lernen.

Mit der Forderung der Bereitstellung eines Möglichkeitsraums, in dem die Lernenden selbst die Entscheidungen über ihren konkreten Lernverlauf als Abbildungsprozess treffen können wird ein hoher didaktischer Anspruch formuliert, sowohl in Hinblick auf die Lernumgebung als auch auf die Lernenden. Wie wird dieser Möglichkeitsraum in der Lernumgebung implementiert? Wie werden Lernende bei der zielgerichteten Navigation unterstützt? Allgemein muss die Lernumgebung die Abbildung einer Vielzahl von Lernwegen ermöglichen, die Abbildung unterschiedlicher Lernstrategien. „Es bedeutet den entscheidenden Paradimawechsel im didaktischen Design, dass in einer professionellen webbasierten Lernumgebung nunmehr alle Parameter des didaktischen Designs variiert und die Variationen als individualisierte Zugänge zum Thema bereitgehalten werden“ (Meder 2006: 59).

14 Diese didaktische Abbildung enthält grundsätzlich ein Moment der Unbestimmtheit, denn wie soll ein vieldimensionaler Gegenstandsbereich in ein lineares Verlaufsschema gebracht werden? Dieses Moment der Unbestimmbarkeit ist reduzierbar durch didaktische Heuristiken, wie z.B. einem deduktiven oder induktiven, einem konstruktiven oder rekonstruktiven, einem synthetischen oder analytischen Vorgehen (vgl. Meder 1995).

Diese *Variation* wird auf Grundlage der didaktischen Ontologie möglich. Die Typisierung durch didaktische Metadaten wird im Folgenden am Beispiel der Wissenseinheiten[15] verdeutlicht. Generell stellen Wissenseinheiten die kleinsten (Informations)Einheiten der Lernumgebung dar und werden in der Web-Didaktik als Antworten auf Fragen verstanden. Auf der obersten Ebene der Wissensarten kann dabei *Orientierungswissen* (know what, know if: Was gibt es überhaupt?) von *Erklärungswissen* (know why: Warum ist etwas so, wie es ist?), *Handlungswissen* (know how: Wie ist etwas anwendbar?) und *Quellenwissen* (know where: Wo finde ich weiteres Wissen?) unterschieden werden.[16] Ein Gegenstandsbereich wird entsprechend der didaktischen Ontologie in Wissenseinheiten aufbereitet (dekontextualisiert). Diese Wissenseinheiten werden thematisch zu Lerneinheiten verknüpft (rekontextualisiert). Die Lerneinheiten stellen dabei einen *Container* für unterschiedliche Wissenseinheiten zum gleichen Thema dar.

Über die Typisierung durch Wissensarten wird ein differenzierter und individualisierter Zugang zu Themen möglich, sowie die Adaption an unterschiedliche Lernende sowie unterschiedliche Lernstrategien[17]. Zentral für diese Typisierung ist die *Granularität* der Wissenseinheit als die Zuordnung *einer* Wissensart zu *einer* Wissenseinheit.[18] Als eine der wichtigsten Maximen der Web-Didaktik bezeichnet demnach Meder (2006: 57) die „Maxime zur Vielfalt des Wissens in Lerneinheiten: Sinn und Zweck der Vielfalt von Wissenseinheiten in einer Lerneinheit besteht darin, alle möglichen Zugänge zum Thema für alle möglichen Lernerinnen bereitzustellen.“[19] Über eine Vielfalt von Wissenseinheiten zu einem Thema (einer Lerneinheit), die jeweils eine Antwort auf eine spezifische Frage darstellen, wird also ein Thema auf vielfältige Weise in der Lernumgebung abgebildet und ermöglicht auf dieser Grundlage für den Lernenden vielfältige zielgerichtete Zugänge zu diesem Thema.

Diese *Vielfalt der Zugänge zum Wissen in der Lernumgebung* ist der zentrale Anknüpfungspunkt der Navigationsanalyse als Analyse *unterschiedlicher empirischer Zugänge zu einem spezifischen Thema,* als Analyse empirischer Navigationsverläufe in der Lernumgebung. Analysiert wird dabei analog zum zweiten Hauptsatz der Didaktik die empirische Abbildung des zeitlich verlaufenden Lernprozesses in den sachlogischen Raum der Bedeutungsbeziehung. Bei der durchgeführten explorativ-heuristischen Studie (vgl. Kap. 9) werden konkret die Navigationssequenzen innerhalb der Lerneinheiten „Maße der zentralen Tendenz“, „Arith-

15 Im Folgenden beziehe ich mich vor allem auf die *rezeptiven* Wissenseinheiten, da diese den Fokus der Navigationsanalyse bilden: Interaktive und kommunikative Wissenseinheiten werden an dieser Stelle nicht weiter ausgeführt.

16 Zur differenzierten Darstellung der Wissensarten, vgl. Meder (2006, 119f.) sowie Anhang 17.11: *Ontologie der rezeptiven Wissensarten in der Web-Didaktik* für einen Überblick.

17 In dieser Arbeit wird der Begriff der *Lernstrategie* bzw. des *Lernstils* verwendet und nicht der Begriff *Lerntypen*, da sich die Navigationsanalyse auf die beobachtbaren Tätigkeiten eines Akteurs bezieht und nicht auf Merkmale des Akteurs, der diese Tätigkeiten vollzieht (vgl. Meder 2006: 212).

18 „Damit das Wissen nachvollziehbar auf verschiedene Lerntypen adaptiert werden kann, enthält jede Wissenseinheit *nur Wissen einer Wissensart* und in der Regel *nur einer spezifischen medialen Darstellung*, sowie nur einer damit verbundenen Kompetenzart“ (Meder 2006: 56; Hervorhebung im Original).

19 Zur Vielfalt didaktischer Modelle, vgl. Flechsig (1983, 1996).

metisches Mittel“, „Median“ und „Modus“ aus dem Bereich der Statistik analysiert.[20] Den Fokus der durchgeführten Navigationsanalyse bildet die Navigation innerhalb der *rezeptiven Lerneinheiten* als Sequenz von Wissenseinheiten. Ausgeblendet bleibt die Nutzung *interaktiver* und auch *kommunikativer Wissenseinheiten*. Diese werden zwar in der Analyse berücksichtigt, soweit sie als Element in der Navigationssequenz enthalten sind, jedoch nicht detailliert in ihrer konkret-inhaltlichen Nutzung.

3.2 Navigation als Autodidaktik

Im folgenden Abschnitt wird aus didaktischer Perspektive der Navigationsprozess in Online-Lernumgebungen näher beschrieben. Dabei wird besonders auf die Navigation als *Autodidaktik* eingegangen. Abschließend wird der Navigationsprozess in der Diskussion um die Selbststeuerung von Lernprozessen verortet.

Allgemein definiert Meder *Navigation* als den konkreten Weg eines Lernenden durch die Wissenseinheiten einer Online-Lernumgebung: „Der Weg, den ein Lerner durch das semantische Netz oder durch die verschiedenen Wissenseinheiten im Mikrobereich – alles in allem durch den Stoff – geht, nennen wir *Lernnavigation* oder auch nur *Navigation.* Den leitenden Gesichtspunkt der Navigation, seine Orientierung an einem sachlogischen, medialen oder didaktischen Prinzip nennen wir *Lernstrategie*. Je nachdem in welchem Bereich oder mit Bezug auf welche Struktur sprechen wir auch von *Makro-Lernstrategie* oder von *Mikro-Lernstrategie*“ (Meder 2006: 62; Hervorhebung um Original). Die Navigationsanalyse untersucht die *Navigationen* unterschiedlicher Lernender als Wege durch das semantische Netz von Wissenseinheiten. Im Vordergrund steht dabei die *Lernstrategie* als Orientierung an sachlogischen Prinzipien und genauer: die *Mikro-Lernstrategie* als *Mikro-Navigation* innerhalb von Lerneinheiten.[21]

Wie kann nun dieser Prozess der Navigation näher gekennzeichnet werden? Von einer im engeren Sinn lernpsychologischen Perspektive grenzt Meder (1995: 61) eine didaktische Perspektive ab: „Bedeutsam ist darüber hinaus die didaktische Seite – nämlich dies, dass der Lernende seinen Lernprozeß selbst gestaltet, d.h. er bestimmt selbst die Metaregeln für die Abbildung seiner Lernzeit in den Bedeutungskomplex der Sachlogik.“ Analog zum zweiten *Hauptsatz der Didaktik* eignet sich der Lernende Wissen an, in dem er seinen zeitlich verlaufenden Lernprozess in den sachlogischen Raum von Bedeutungsbeziehungen abbildet, also auf Grundlage seines konkreten Navigationsverlaufs ein mentales Modell des Gegenstandsbereichs entwi-

20 vgl. Abbildungen der Wissenseinheiten im Anhang („Maße der zentralen Tendenz“, Kap. 17.12; „Arithmetisches Mittel“, Kap. 17.13; „Median“, Kap. 17.14; „Modus“, Kap. 17.15).

21 Prozesse der Makronavigation als Navigation zwischen Lerneinheiten werden in dieser Arbeit nicht weiter verfolgt, sind aber grundsätzlich auf gleicher theoretischer Grundlage und mit gleicher Methodologie möglich.

ckelt. Im Rahmen der Web-Didaktik wird diese *Selbstgestaltung* des Lernprozesses als „Autodidaktik“ (Meder 2006: 65) bezeichnet, als Steuerung des didaktischen Prozesses durch die Lernenden.

Diese *Selbstdidaktik* des Übersetzens von Zeitgestalten in Raumgestalten lässt sich als entgegengesetzer Prozess der *Didaktik* als Abbildung von Raumgestalten in Zeitgestalten verstehen. *Autodidaktik* bedeutet demnach für den Lernenden, dass er selbst die didaktischen Entscheidungen der Abbildung treffen kann - und treffen muss. Der Lernende muss also über Lernstrategien in Form von *Metaregeln* für seine Entscheidungen der Abbildung verfügen. Insofern ermöglicht die Lernumgebung vielfältige Navigationsweisen und erfordert und fördert gleichzeitig die Reflexion des eigenen Navigationsprozesses.

Im Rahmen der Exploration von Online-Lernumgebungen muss der Lernende diese Abbildung selbst leisten. Er geht dabei von der Lernzeit aus und bildet diese in die Sachlogik ab. Da es sich bei der Selbstdidaktik um eine anspruchsvolle und komplexe Tätigkeit handelt (vgl. Iske 2002), werden die Lernenden durch die didaktische Struktur der Lernumgebung bei diesem Abbildungs- bzw. Lernprozess unterstützt. Auf der Seite der Lernenden erfordert dies eine „autodidaktische Kompetenz“ (Meder 2006: 33) als ein Wissen um das eigene Lernen.

In der hier vorliegenden Navigationsanalyse bildet also weniger die *angeleitete* Navigation in einer Online-Lernumgebung den Schwerpunkt, wie sie z.B. oft in Form von *Kursen* (als *guided tour*, vgl. Iske 2002) implementiert ist, sondern vielmehr die *explorative* Navigation als entdeckend-lernendes Navigieren.

Der Kurs als *angeleitete Navigation* entspricht dabei der Abbildung des Gegenstandsbereichs in einen spezifischen zeitlichen Verlauf, wie er in der konkreten Anordnung der Wissenseinheiten des Kurses zum Ausdruck kommt. *Angeleitet* ist diese Form der Navigation, da die Reihenfolge der Wissenseinheiten des Kurses von einem Autor im Vorfeld definiert wurde (vgl. erste Hauptsatz der Didaktik): Aus einer Vielzahl möglicher Reihenfolgen (Abbildungen) hat der Autor den Kurs als eine konkrete Reihenfolge in die Lernumgebung implementiert. Die Nutzenden der Lernumgebung haben dann grundsätzlich die Möglichkeit, neben der freien Exploration der Lernumgebung diesem Kurs als definierter Anordnung von Wissenseinheiten zu folgen.

Hinsichtlich der Kurse ist auf eine *Besonderheit der Implementierung* in der analysierten Lernumgebung *Lerndorf* hinzuweisen, die für die Analyse und Interpretation der Navigationssequenzen von Bedeutung ist (vgl. Abb. 4): Entscheidet sich die Nutzerin für einen Kurs, ändert sich die Oberfläche der Lernumgebung. Es wird ein „Player“ angezeigt, mit dem sie innerhalb der Abfolge des Kurses vor- und zurück gehen kann. Darüber hinaus werden in der linken Navigationsleiste (Themenbaum) nur die Lerneinheiten angezeigt, die im Kurs enthalten sind (bei der freien Navigation werden alphabetisch sortiert *alle* Lerneinheiten der Lernumgebung im Themenbaum angezeigt). Obwohl sich die Nutzerin für die Kursnavigation entschieden hat, kann sie durch Auswählen von Lerneinheiten im Themenbaum die vom Kursautor festegelegte Reihenfolge verlassen und innerhalb der Kurseinheiten *frei*, d.h. in selbst gewählter Reihenfolge navigieren. Unterbricht

eine Nutzerin die Navigation innerhalb des Kurses indem sie den Kurs verlässt, wird beim Wiedereinstieg in den Kurs genau die Wissenseinheit angezeigt, die zuletzt aufgerufen wurde.

Weiter oben wurde bereits darauf hingewiesen, dass der Prozess der Mikro-Navigation den Gegenstand der vorliegenden Arbeit bildet. Zentral für die Analyse der Mikro-Navigation ist das Konzept der *didaktischen Stationen* (vgl. Meder 2006: 208). Diese bezeichnen Orte innerhalb der Lernumgebung, an denen die Nutzenden eine Entscheidung über das weitere Vorgehen treffen müssen. Eine solche didaktische Station innerhalb einer Lerneinheit stellt die Wissenseinheit *Orientierung* dar. Wählt der Nutzer bei freier Exploration der Lernumgebung eine konkrete Lerneinheit aus, so wird die zugehörige Wissenseinheit *Orientierung* angezeigt: *Orientierungswissen* stellt also den *Startpunkt* der jeweiligen Lerneinheit dar. Sie ist im Rahmen der Navigationsanalyse somit der zentrale Ausgangspunkt der Navigationsverläufe. Hauptgegenstand der Navigationsanalyse ist dann der weitere Navigationsverlauf innerhalb dieser Lerneinheit, d.h. als Navigationsverlauf zwischen Wissenseinheiten dieser Lerneinheit (Mikronavigation).

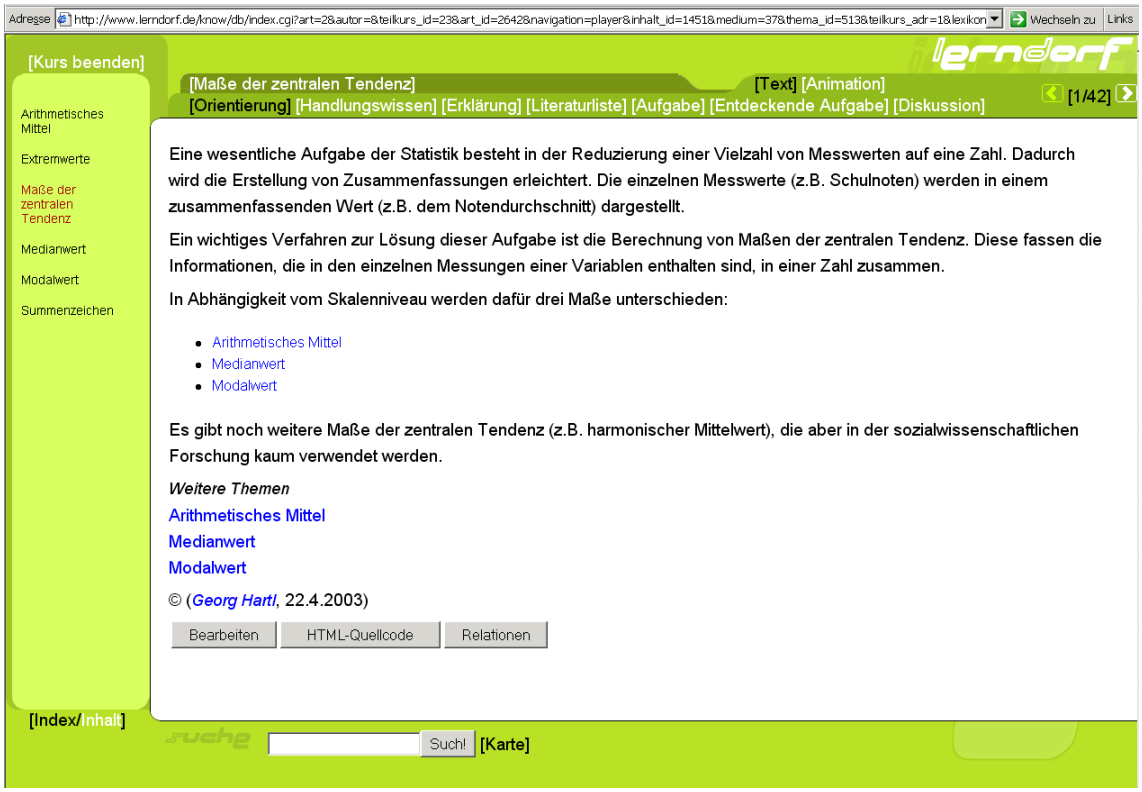

Abbildung 4: Kurs: "Statistik - Maße der zentralen Tendenz"

Wie ist dieser Prozess der Selbstdidaktik im weiteren Kontext der Diskussion um *selbstgesteuertes Lernen* zu verorten? Die Suche nach einem neuem Paradigma des Lernens bzw. nach einem neuem Bildungskonzept, das die Aktivität des Lerners betont und sich bewusst vom traditionellen schulischen Lehrkonzept unterscheidet führt über die Konzeption einer „lifelong education“ (Lengrand 1972, 1975, 1987) über das „lifelong learning“ (Delors 1998, 2004) zum „selfdirected learning“ bzw. „selbstgesteuerten Lernen“ (Doh-

men 1998). In der Literatur finden sich neben dem grundlegenden Begriff des „selbstgesteuerten Lernens“ zahlreiche weitere Begriffe, die in ähnlicher Zielrichtung oft synonym verwendet werden.

Ausdrücklich sei an dieser Stelle darauf hingewiesen, dass die Vorsilbe „Selbst-“ im Sinne von „aus sich selbst“ zu verstehen ist, und nicht als „von selbst“: die Lernenden müssen also selbst tätig, selbst aktiv werden. Aus dem selbstgesteuerten Lernen resultieren veränderte Anforderungen an das Lernen und dieser Paradigmenwechsel wird in der Literatur gekennzeichnet als Verschiebung des Fokus und als Übergang

- vom Lehren zum Lernen,[22]
- vom lehrerorientierten /-zentrierten zum lernerorientierten /-zentrierten Unterricht,
- vom fremd- zum selbstgesteuerten Lernen,
- vom schulischen Lernen zum lebenslangen Lernen,
- vom reaktiven zum aktiven Lernen,
- vom darbietenden zum erarbeitenden, erkundenden, problemlösenden Unterricht,
- vom didaktischen Dreieck zum Lernarrangement (Topologie),
- von der Vermittlungsdidaktik zur Arrangementdidaktik,
- vom Darbieten von (Erkenntnis)-Produkten zum Anregung von (Erkenntnis)-Prozessen,
- von der Kenntnisvermittlung zur Erkenntnisvermittlung,
- von deklarativem Wissen zu prozeduralem Wissen,
- von angeleitetem Lernen zu selbsttätigem Lernen.

Gemeinsam ist diesen Beschreibungen, dass die Lernenden zum Ausgangspunkt und Zentrum des Lehr-Lern-Prozesses werden, mit einem höheren Grad an Aktivität und mehr Verantwortung als in Unterrichtsformen, in denen Methoden der Vermittlung und Darbietung überwiegen. Ausgangspunkt für das Konzept des selbstgesteuerten Lernens ist dabei die elementare Einsicht, „dass Lernen ein Prozess ist, der vom Lernenden selbst realisiert werden muss“ (Bönsch 2000: 186).

Wie Bönsch (2000) macht auch Brinkmann (2000) deutlich, dass beim selbstgesteuerten Lernen graduelle Unterschiede bestehen, die sich zwischen den Polen der vollständigen Selbststeuerung von Lernprozessen und der Selbststeuerung in Form eigener Ziel- und Richtungsentscheidungen bewegen. Grundsätzlich geht es bei dem Paradigma des selbstgesteuerten Lernens „im Kern im die eigene Ziel- und Richtungsentscheidung der Lernenden, nicht in jedem Fall um die Selbstorganisation von Lernprozessen. Auch institutionelle Lernhilfen und Angebote können in ein zielgerichtetes individuelles Lernen eingebunden sein“ (Brinkmann 2000: 54). Damit verbunden ist eine Lernkultur, in der die Lernenden stärker als bisher Lernziele, Lerninhalte und Lernwege gestalten und bestimmen. Als Eckpunkte der Selbststeuerung von Lernprozessen nennt Brinkmann (2000: 54-55) das Ziel (wohin, wofür?), die Inhalte (was?), den Lernweg (wie, auf welche Wei-

22 „Wenn wir also von 'Lernformen' sprechen, ist die 'Lehrform' stets mitgedacht. Sie steht aber nicht mehr im Zentrum, denn die Steuerung des Verhältnisses von Lehren und Lernen wird zunehmend in die Verantwortung der Lernenden gelegt. Sie beinhaltet ein spezifisches Arrangement von individuellen und medialen Lerntechniken, Lehrer- und Lernerrollen sowie Zugänge zu Lerninhalten und raum-zeitlichen Bedingungen“ (Brinkmann 2000: 35-36).

se, mit wem, mit welchen Hilfsmitteln?), die Überprüfung des Lernerfolges (wie gut?) und die Lernregulierung (wann, wo, wie lange?).

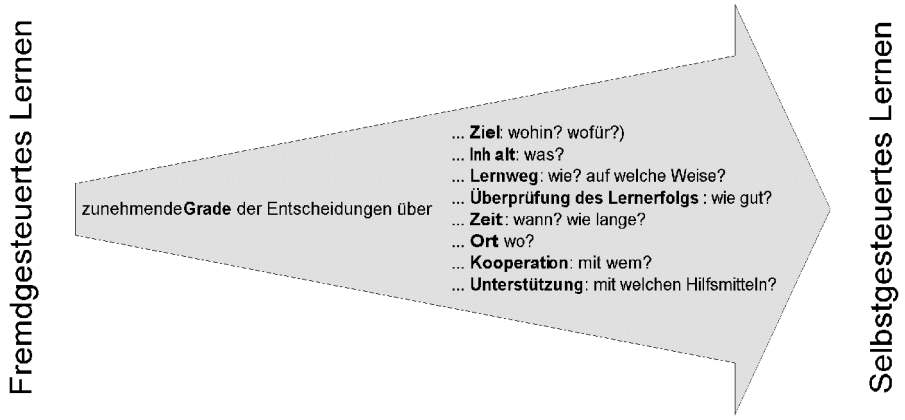

Abbildung 5: Fremdgesteuertes und selbstgesteuertes Lernen

Auf ein grundlegendes Problem der Theorie des selbstgesteuerten Lernens macht Swertz (2004a) aufmerksam: Die Diskussion um Entscheidungsfelder selbstgesteuerten Lernens (vgl. Tabelle 1 sowie Abbildung 5) entspricht den didaktischen Entscheidungsfeldern der Berliner Didaktik (vgl. Heimann 1976). Diese wurden jedoch in der Absicht entwickelt, Lehrende bei der Planung von Unterricht, also der Fremdsteuerung von Lernprozessen, zu unterstützen. Swertz (2004a) weist darauf hin, dass es sich bei Selbststeuerung und Fremdsteuerung nicht um kontradiktorische Konzepte handelt, zwischen denen zu entscheiden ist, sondern um konträre Konzepte, die es zu vermitteln gilt, also um notwendige Momente von Bildungsprozessen. „Schon die Überlegung, dass die Ermöglichung einer Selbststeuerung oder Selbstevaluation eine Fremdsteuerung darstellt zeigt, dass Selbst- und Fremdsteuerung nicht kontradiktorisch, sondern konträr gegenüber gestellt werden müssen." (Swertz 2004a: 177) So ist es im Rahmen selbstgesteuerten Lernens in einer Online-Lernumgebung möglich, sich – selbstgesteuert - für die Bearbeitung eines – fremdgesteuerten - Kurses zu entscheiden. Die Entscheidungsfelder in Tabelle 1 sowie Abbildung 5 verdeutlichen, dass es sich bei der Selbststeuerung von Lernprozessen um eine komplexe und anspruchsvolle Tätigkeit handelt.

Während in der Diskussion um E-Learning vor allem die Aspekte der räumlichen und zeitlichen Unabhängigkeit des Lernens im Vordergrund stehen, zielt die Web-Didaktik bzw. *Autodidaktik* darüber hinaus auf die Abbildung des zeitlichen Lernprozesses in die sachlogische Bedeutungsbeziehung. Damit steht neben den Entscheidungen über die Auswahl des Inhaltes vor allem die Entscheidung über den Lern*weg* im Vordergrund. Eine Analyse des genauen Lernwegs als Lernprozess im engeren Sinn stellt in der Diskussion um E-Learning gegenwärtig einen blinden Fleck dar, da dieser Aneignungs- bzw. Lernprozess in den gängigen E-Learning Plattformen in der Regel weder beobachtbar noch analysierbar ist.

Auf Grundlage der oben beschriebenen didaktischen Ontologie und besonders der modularisierten Wissenseinheiten sind in der analysierten Online-Lernumgebung theoretisch und praktisch eine Vielzahl und Vielfalt von Lernwegen möglich, die jeweils von den Entscheidungen der Lernenden abhängen. Diese vielfältigen Lernwege sind innerhalb der Lernumgebung beobachtbar, dokumentierbar und analysierbar.

Orientierung des Lerngeschehens	Lerner-	Lehrerzentrierung
Aktivitätsgrad des Lernenden	Agierender-	Konsumierender Lerner
Zeitliche Flexibilität	Flexible-	Gebundene Lernzeiten
Räumliche Flexibilität	Variable-	Feste Lernorte
Entscheidungsfreiheit über Lernziele	Lernzielautonomie-	Vorgegebene Lernziele
Entscheidungsfreiheit über Lerninhalte	Frei wählbare-	Vorgegebene Inhalte
Überprüfung des Lernerfolgs	Selbstkontrolle	Fremdkontrolle

Tabelle 1: Eckpunkte der Selbststeuerung (vgl. Gnahs (1998: 28f.), in: Brinkmann (2000: 40).

3.3 Zusammenfassung

Zusammenfassend können folgende Vorteile der Verwendung einer auf der Web-Didaktik basierenden Lernumgebung für die Navigationsanalyse festgehalten werden: die Analyse differenzierter Navigationsverläufe; die Rekonstruktion und Analyse von Prozessen der Autodidaktik als Abbildung von Zeitgestalten in Raumgestalten; die inhaltliche Interpretation der Sequenzen auf Grundlage der didaktischen Ontologie sowie der Vergleich der Navigationssequenzen in unterschiedlichen Bereichen (Lerneinheiten) aufgrund der Typisierung der jeweiligen Wissenseinheiten durch Metadaten (vgl. Meder 2006).

Grundsätzlich ist zu betonen, dass die Methodologie der Navigationsanalyse und vor allem das Optimal-Matching Verfahren nicht auf eine auf der Web-Didaktik basierende Lernumgebung beschränkt ist. Eine Analyse der Navigationsverläufe als Sequenzen ist selbstverständlich auch in anderen Lernumgebungen möglich. Jedoch ermöglicht gerade die *didaktische Ontologie* ein besonderes Potenzial aufgrund der *Modularisierung* und *Granularität* von Wissenseinheiten: jede Wissenseinheit enthält genau eine Wissensart. Dieser besondere Grad an *Differenzierung* kann in nicht-metadatenbasierten Lernumgebungen nicht erreicht

werden und ist auch in Lernumgebungen, die auf Metadatenstandards wie LOM, SCORM oder IMS-LD basieren[23] in dieser Form bisher nicht erreichbar.

So ist in den als „Document Management Center“ (vgl. Kap. 2) kritisierten Plattformen ein solcher Differenzierungsgrad der Analyse und Interpretation von Aneignungsprozessen nicht möglich, da der eigentliche Lernprozess aus diesen Plattformen ausgegliedert und auf PDF-Dokumente übertragen wird. Die Aneignung von Wissen als Abbildung von Lernzeit in die Sachlogik ist dabei weder beobachtbar noch analysierbar. So kann in der Plattform „Blackboard“[24] zwar der Navigationsverlauf anhand von Logfiles rekonstruiert werden und es wird beispielsweise erkennbar, dass eine Nutzerin ein konkretes PDF-Dokument auf ihre Festplatte kopiert hat. Der eigentliche Aneignungsprozess findet dann aber in der Auseinandersetzung mit diesem PDF-Dokument statt, ist nicht mehr Teil der Plattform und kann als solcher nicht rekonstruiert werden (Liest die Lernende das PDF-Dokument oder nicht? Liest die Lernende das PDF-Dokument komplett oder auszugsweise? Welche Informationen liest sie, welche überspringt sie?). Der oben beschriebene Prozess der Autodidaktik ist dabei nicht rekonstruierbar. Eine vergleichbare Schwierigkeit besteht bei der Analyse von Navigationsprozessen im WWW: In der Regel kann die besuchte Seite nicht eindeutig sachlogisch klassifiziert werden, da vielfältige und unterschiedliche Informationen (Wissensarten) auf einer angezeigten Seite enthalten sind. Es bleibt also unklar, ob überhaupt und wenn ja welche Informationen auf der Seite konkret genutzt wurden (vgl. Kap. 4.2; *Web-Mining*).

Im Rahmen der Web-Didaktik ist grundsätzlich über die eindeutige Zuordnung einer Identifikationsnummer zu jeder Wissenseinheit der Inhalt der besuchten Wissenseinheiten rekonstruierbar. Die Typisierung von Wissenseinheiten auf Grundlage der didaktischen Ontologie ermöglicht über diese inhaltliche Interpretation hinaus eine Analyse auf der Abstraktionsebene der didaktischen Metadaten. Da die didaktische Ontologie den Kern der Lernumgebung darstellt und die Navigationsverläufe in dieser didaktischen Ontologie abgebildet werden, kann der Navigationsprozess insgesamt aus didaktischer Perspektive interpretiert werden, z.B. auf der Folie didaktischer Heuristiken.

Zusätzlich können aufgrund der didaktischen Metadaten auch Navigationsverläufe unterschiedlicher Bereiche (d.h. unterschiedlicher Lerneinheiten) als Abfolge von Wissenseinheiten oder Wissensarten interpretiert und miteinander verglichen werden. Dabei abstrahiert die Interpretation von dem konkreten Inhalt der Wissenseinheit und analysiert den Navigationsverlauf auf der Ebene der didaktischen Metadaten.

Der dargestellte Grad an Differenziertheit ist erforderlich, wenn Aneignungsprozesse von Lernenden *während* des E-Learning in Form von Navigationsverläufen den zentralen Fokus der Analyse bilden. Unter dieser Voraussetzung können Prozesse der Aneignung hypertextueller Online-Lernumgebungen als das lineare

23 Vgl. „Learning Object Metadata“ (LOM), <http://ltsc.ieee.org/wg12/>, (28.08.2006); „Sharable Content Object Reference Model“ (SCORM), <http://www.adlnet.gov/index.cfm>, (28.08.2006); „Instructional Management System -Learning Design“ (IMS-LD), <http://www.imsglobal.org>, (28.08.2006).

24 Vgl. Blackboard Inc., <www.blackboard.com>, (28.08.2006).

Entfalten eines nicht-linearen Hypertextes (Kuhlen 1991: 33, vgl. Kap. 2: 7) in den Verlauf der Bearbeitungszeit analysiert werden. Genau dieses *lineare Entfalten* (Navigationsverlauf) eines *nicht-linearen Gegenstandsbereichs* (Hypertext) kann analog zu Meder und Hönigswald als Autodidaktik interpretiert werden, als Abbildung der Zeit der Navigation in die Sachlogik des Gegenstandsbereichs. Damit geht die Navigationsanalyse weit über eine rein formale Analyse aggregierter Logdaten hinaus.

4 Kontext der Navigationsanalyse

In diesem Kapitel wird am Beispiel der *Analyse von Logfiles* und des *Web-Mining* der Forschungskontext der Navigationsanalyse dargestellt.

Zur Abgrenzung wird die Logfile-Analyse als Analyse *aggregierter* Logdaten näher ausgeführt und im nächsten Kapitel der Analyse *sequenzierter* Logdaten gegenübergestellt. Aggregierte wie auch sequenzierte Logdaten basieren auf den Zugriffen von Nutzenden auf Online-Umgebungen, unterscheiden sich jedoch hinsichtlich der Datenaufbereitung und der darauf aufbauenden Analyse.[25] Die Datenaufbereitung als ein der Optimal-Matching Analyse vorgeschalteter Schritt wird in Kapitel 10 dargestellt, die Methode der Sequenzanalyse mittels Optimal-Matching in Kapitel 6.

Am Beispiel des *Web-Mining* wird ein Forschungsfeld dargestellt, das wie die Navigationsanalyse die Handlungen von Nutzenden in Online-Umgebungen analysiert, jedoch mit unterschiedlichem Schwerpunkt, auf unterschiedlicher methodologischer Grundlage und mit unterschiedlicher Zielsetzung.

4.1 Analyse aggregierter Logdaten

In diesem Abschnitt wird die Analyse *aggregierter* Logdaten dargestellt, um darauf aufbauend in den folgenden Kapiteln das Potenzial der *sequenzierten* Analyse deutlich zu machen.[26]

Allgemein ist festzuhalten, dass die Analyse *aggregierter* Logdaten das am weitesten verbreitete Verfahren zur Analyse der Nutzung von Online-Umgebungen darstellt. Viele Internet-Provider stellen ihren Kunden die Analyse der Logdaten in Form *aggregierter* Nutzungsdaten zur Verfügung. *Aggregiert* bedeutet in diesem Zusammenhang, dass die in den Logfiles enthaltenen Informationen zusammengefasst werden und darauf aufbauend *durchschnittliche Kennzahlen* berechnet werden (z.B. durchschnittliche Nutzungsdauer, durchschnittliche Anzahl der besuchten Seiten, sowie Minimum- und Maximum-Werte). Im Vordergrund steht dabei die deskriptiv-statistische Analyse der Logfile-Daten mit der grundlegenden Orientierung am Querschnittsdesign der Datenerhebung und Datenanalyse (vgl. Trautner 1992).

25 In dieser Arbeit wird der Begriff der „Logfile-Analyse“ in Anlehnung an den üblichen Sprachgebrauch zur Bezeichnung der Analyse *aggregierter* Logfile-Daten verwendet. Für die Bezeichnung der Analyse *sequenzierter* Logfile-Daten wird der Begriff der „Sequenzdatenanalyse“ verwendet.

26 Für einen kompakten Überblick über die Analyse aggregierter Logdaten, vgl. Priemer (2004).

Die grundlegenden Analyseeinheiten der *Sequenzdatenanalyse* (und der Optimal-Matching Analyse) sind im Gegensatz zur Analyse *aggregierter* Daten jedoch *sequenzierte* Daten als *verlaufsbezogene* Daten. Diese stellen die Navigationsverläufe von Nutzenden in einer Online-Umgebung dar: der Verlauf der besuchten Internetseiten wird dabei als Sequenz dokumentiert und analysiert. Diese sequenzierten Daten wurden in Kapitel 2.1.2 bereits als „elektronische Prozessdaten" (Bergmann / Meier 2000: 431) beschrieben. In den sequenzierten Daten sind also Informationen über den zeitlichen *Verlauf* enthalten, die in den aggregierten Daten nicht mehr enthalten sind. Allgemeines Kennzeichen von Verlaufsdaten ist die Orientierung am Längsschnittdesign der Datenerhebung und Datenanalyse, d.h. die Daten werden *wiederholt* in definierten zeitlichen Intervallen beim *gleichen* Individuum erhoben.[27] Während Sequenzdaten die Analyse des Verlaufs bzw. der Entwicklung des Navigationsprozesses ermöglichen, sind diese Analysen auf der Grundlage aggregierter Daten aus methodologischen Gründen nicht möglich.

Sozialwissenschaftliche Forschungen zur Analyse der Internetnutzung basieren überwiegend auf der Analyse aggregierter Logdaten: Bei der Analyse stehen deskriptive und inferenzstatistische Methoden im Vordergrund.[28] Bei diesem Forschungsdesign sind jedoch die *Prozesse* der Nutzung selbst – z.B. der konkrete zeitliche Verlauf der Navigation als *Sequenz* - nicht Gegenstand der Analyse.

Programme zur Logfile-Analyse rekonstruieren aus den Logdaten in der Regel jedoch nicht den Navigationsverlauf als sequenzierte Daten, sondern aggregieren diese Daten. Dieser Ansatz kommt deutlich bei der Funktionalität von Standard-Software zur Analyse von Logdaten zum Ausdruck: Wie lange und wie häufig werden Seiten besucht? Bei welchen Seiten steigen die Nutzer ein, bei welchen aus? Wie häufig kehren Nutzerinnen zu der Online-Umgebung zurück? Mit welchen Browsern greifen Nutzer auf das Angebot zu? Zu welchen Tagen, zu welchen Uhrzeiten greifen Nutzerinnen auf das Angebot zu? Dies entspricht der Berechnung von *durchschnittlichen Kennzahlen* bzw. *Zustandsverteilungen,* um die Nutzung von Online-Plattformen zu analysieren. Fokussiert wird dabei der „Wandel von Mischungsverhältnissen" in den aggregierten Daten (Baur 2005: 167).

Dazu werden in der Regel Kennzahlen wie „Hits", „Page Views", „Visitors" und „Session" verwendet. Die Kennzahl „Hits" bezieht sich auf die Anzahl aller Objekte, die vom Server über das Internet zum Nutzer gesendet werden (z.B. Bilder, Grafiken, Animationen u.ä.) und die der Browser des Nutzers zum Aufbau und zur Anzeige der angeforderten Seite benötigt. Die Anzahl der „Hits" entspricht einer rein formalen Beschreibung und sagt nichts über die Qualität oder über die inhaltliche Nutzung der Objekte aus. Die Anzahl der „Hits" variiert je nach Aufbau und Struktur der Seite: Internetseiten, die aus vielen einzelnen Objekten bestehen erzeugen somit mehr „Hits" als Seiten, die aus wenigen Objekten aufgebaut sind. Die Nutzung unter-

27 Bei der Navigationsanalyse liegt das zeitliche *Intervall* der Datenerhebung bei null, d.h. jede Handlung in Form eines Klicks wird mit der Information über den genauen Zeitpunkt der Handlung in den Logfiles aufgezeichnet (vgl. Kap. 2, *Methodologische Grundlagen der Navigationsanalyse*).

28 Vgl. ARD-ZDF Online Studie (van Eimeren / Frees 2005); JIM-Studie 2005 (Feierabend / Rathgeb 2005), (N)onliner-Atlas (Möller 2006), UK Children go Online (Livingstone / Bober 2005).

schiedlicher Internet-Angebote kann anhand der „Hits“ nicht ohne weiteres verglichen werden. Insbesondere sind „mehr Hits“ nicht mit „mehr Nutzung“ gleichzusetzen.

Die Kennzahl „Page Views“ bezieht sich auf die vom Nutzer betrachteten *Seiten*. Es wird dabei nur die *Anzahl* der Seiten berücksichtigt, die vom Nutzer angefordert wurden. Technisch gesehen beziehen sich die „Page Views“ beispielsweise auf übermittelte HTML-, XML- oder PHP-Seiten. Objekte wie Bilder und Animationen werden dabei nicht berücksichtigt. Der Aufbau und die Struktur der Seiten spielt bei der Kennzahl „Page Views“ keine Rolle.

Die Kennzahl „Besucher“ („Visitors“) bezieht sich auf die Anzahl der Zugriffe unterschiedlicher Computer (genauer: auf den Zugriff unterschiedlicher IPs) auf den betreffenden Server, unabhängig von der Anzahl der dabei entstandenen „Hits“ oder „Page Views“. Unterschiedliche IPs werden dabei als unterschiedliche Besucher interpretiert. Diese Analysen sind jedoch mit relativen Ungenauigkeiten behaftet (variable IP-Adressen, Cache, Proxi-Server). Darüber hinaus ist nicht jeder in den Logfiles aufgezeichnete „Nutzer“ auch eine „reale“ Person, sondern kann auch auf automatisierte Zugriffe von Suchmaschinen, Robots oder Spider zurückzuführen sein, die im Vorfeld einer Logfile-Analyse zu identifizieren und zu filtern sind.

Die Kennzahl „Session“ wird in Anlehnung an die Kennzahl „Besucher“ errechnet. Der entscheidende Unterschied besteht in der Berücksichtigung des Faktors Zeit. Wenn ein Besucher eine Seite aufruft und wieder verlässt, um zu einem späteren Zeitpunkt zurückzukehren, wird dies als *zwei* Sessions interpretiert (im Gegensatz zur Interpretation als *ein* Besucher). Das zeitliche Intervall zur Interpretation als neuer Session zwischen Verlassen der Seite und Wiederkehr wird in der Praxis der Logfile-Analyse in der Regel mit 30 Minuten definiert (vgl. Srivastava et al. 2000: 3).

Der Prozess der *Nutzung* der Webseite kommt bei diesen auf aggregierten Daten beruhenden Kennzahlen lediglich als *Session* bzw. *durchschnittliche Sessiondauer* in den Blick. Die Darstellung der Navigationsverläufe einzelner Nutzerinnen verbleibt im Status der Darstellung nebeneinander stehender, unverbundener Fälle, die mit Programmen zur Logfile-Analyse nicht weitergehend systematisch analysiert werden können.[29]

Die Sequenzanalyse im Rahmen der Navigationsanalyse setzt genau an diesem Punkt an: Der Navigationsverlauf als Sequenz wird zum Ausgangspunkt und zur Analyseeinheit, ohne jedoch wie bei Standard-Logfile-Analyse auf die Aggregation der Logfile-Daten und einer darauf aufbauenden Analyse zurückzugreifen.

Das methodische Repertoire zur quantitativen Analyse von *aggregierten Querschnittsdaten* mit Hilfe deskriptiver und inferenzstatistischer Verfahren wird an dieser Stelle als bekannt vorausgesetzt. Doch wie können *Sequenzdaten* als *Längsschnittdaten* analysiert werden? Was bedeutet „Ähnlichkeit“ in Bezug auf Sequenzen? Wie kann Ähnlichkeit bzw. Distanz zwischen Sequenzen festgestellt werden? Welche Möglichkei-

29 Für einen Überblick über die Funktionalität unterschiedlicher Programme zur Logfile-Analyse, vgl. Kurzidim / Engler (2002), Baketaric / Strübel (2004).

ten gibt es, Sequenzen aufgrund von „Ähnlichkeit“ zu kategorisieren? Diese Fragen stellen einen zentralen Ausgangspunkt der vorliegenden Arbeit dar und werden in Kapitel 5 näher ausgeführt.

4.2 Web-Mining

In diesem Abschnitt wird das Forschungsfeld des *Web-Mining* als Teilgebiet des *Data-Mining* dargestellt. Gemeinsamer Ausgangspunkt des *Web-Mining* wie der Navigationsanalyse ist das Verhalten von Nutzenden in Online-Umgebungen, der Gegenstandsbereich unterscheidet sich jedoch deutlich: *Web-Mining* fokussiert den Bereich des *E-Commerce*, die Navigationsanalyse den Bereich des *E-Learning*. Unterschiede bestehen darüber hinaus im methodischen Vorgehen.

Allgemein beschäftigt sich das *Data-Mining* mit der explorativen Analyse großer Datenbestände mit dem Ziel, relevante Informationen in Form von Mustern, Regelmäßigkeiten und Strukturen zu identifizieren und zu extrahieren (vgl. Fayyad et al. 1994; Chen / Han / Yu 1996, Cooley 2000, Mobasher / Liu / Masand / Nasraoui 2004). Die explorative Ausrichtung der Analyse kommt im Begriff des „Mining“ als „Schürfen“ und „Graben“ zum Ausdruck.

Generell fokussiert *Web-Mining* die Analyse von Daten, die im Internet entstehen. In Abhängigkeit dieses Entstehungskontextes gliedert sich das *Web-Mining* in die Bereiche der *inhaltlichen* Analyse (*content mining)*; der *strukturellen* Analyse (*structure mining)* und der Analyse der *Nutzung* von Online-Plattformen bzw. Webseiten (*usage mining)*. Im Bereich des *usage mining* werden sowohl die Nutzungsdaten (*usage*) in Form von Logdaten (IP-Adressen, Referrer, Datum, Objekte) analysiert, als auch darüber hinaus gehende Nutzerdaten (*user profile*s) in Form von Kunden- und Registrierungsdaten (*customer profile information*) (vgl. Srivastava et al. 2000). Grundsätzlich können dem Web-Usage Mining unterschiedliche Datenquellen zu Grunde gelegt werden: Diese reichen von der Analyse einzelner Nutzer in einer Online-Umgebung (*single-user, single-site)* bis zur Analyse vieler Nutzer auf unterschiedlichen Seiten *(multi-user, multi-site).*

Den historischen Hintergrund des *Web-Mining* bildet die technische Möglichkeit der genauen Rekonstruktion des Nutzungsverhaltens in Online-Umgebungen, die bis zum einzelnen Mausklick reicht. Diese neuartigen technischen Verfahren beschreibt Srivastava et al. (1999: 1) aus der Perspektive des E-Commerce als Revolution: „Specifically, e-commerce activity that involves the end user is undergoing a significant revolution. The ability to track users' browsing behavior down to individual mouse clicks has brought the vendor and end customer closer than ever before“. Aufbauend auf der Analyse der so entstehenden Daten wird ein besonderes Potential vor allem für die Personalisierung von E-Commerce Angeboten abgeleitet. Den Ablauf

des Web-Usage Mining Prozesses unterteilt Srivastava et al. (1999: 1) in drei Phasen: Datenaufbereitung (*preprocessing),* Exploration von Mustern *(pattern discovery)* und Musteranalyse (*pattern analysis).*

Die Datenaufbereitung beruht darauf, die Daten in eine für die weitere Analyse erforderliche Abstraktionsebene zu transformieren: „Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery" (Srivastava et al. 2000: 3). Diese Abstraktionsstufen bestehen in Bezug auf das Web-Usage Mining beispielsweise in der Identifizierung von Nutzern (*user*), der von den Nutzenden aufgerufenen Seiten auf einem oder mehreren Servern (*user session*) oder der Episoden (*episode*) aufgerufener Seiten als Teilmenge einer *user session*.[30] In Bezug auf die inhaltlichen Analyse (content mining) besteht dieser Prozess der Datenaufbereitung und Abstraktion beispielsweise in der Klassifikation oder in der Clusterung von Inhalten von Online-Umgebungen.

Die beim *Web-Mining* verwendeten Methoden zur Datenaufbereitung und Datenexploration sind vielfältig und stammen aus den unterschiedlichsten wissenschaftlichen Bereichen. „Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition" (Srivastava et al. 2000: 5). Zum Repertoire des *Web-Mining* gehört eine Vielzahl deskriptiver, inferenzstatistischer und multivariater Verfahren: von Hauptkomponentenanalysen, Faktorenanalysen, Clusteranalysen, Regressionsanalysen, Assoziations- und Korrelationsanalysen, Multidimensionaler Skalierung und Klassifikation bis zu künstlichen neuronalen Netzen, Hidden Markov Models und Baye'schen Netzen (vgl. Fayyad / Piatetsky-Shapiro / Smyth 1996; Zaïane / Xin / Han 1998; Cooley / Mobasher / Srivastava 1997; Cooley / Tan / Srivastava 1999; Berendt 2001; Berendt / Spiliopoulou 2002; Cadez / Heckerman / Smyth / White 2000; Chakrabarti 2000; Chakrabarti 2003; Mobasher / Dai / Luo / Nakagawa 2002; Oyanagi / Kubota / Nakase 2002; Ypma / Heskes 2002; Chen / Sun / Zaïane / Goebel 2004; Mobasher / Liu / Masand / Nasraoui 2004, Hooker / Finkelman 2004; Lu / Dunham / Meng 2005).

Bekannt ist vor allem die *Warenkorbanalyse* als Beispiel einer Assoziationsanalyse. Dabei steht die Analyse von Nutzungsinformationen (usage mining) in Hinblick auf gemeinsame Interessen der Nutzenden im Vordergrund. Auf Grundlage dieser *Warenkorbanalyse* werden Kaufempfehlungen abgeleitet, wie dies z.B. bei *amazon.de* als dynamisches Empfehlungssystem implementiert ist:[31] „Kunden, die diesen Artikel gekauft haben, kauften auch:", „Kunden, die diesen Artikel angesehen haben, haben auch angesehen:", „Unser Vorschlag: Kaufen Sie jetzt diesen Artikel zusammen mit".

Mit Blick speziell auf die Analyse von *Sequenzen* ist in Bezug auf das Web-Mining festzuhalten, dass Verfahren der Ereignisanalyse auf der Grundlage von Markov-Ketten den methodischen Schwerpunkt bilden.[32]

30 Zur Definition grundlegender Begriffe, vgl. „w*eb usage characterization activity"*, <http:www.w3c.org/WCA>, (28.08.2006).

31 Dieses dynamische Empfehlungssystem berücksichtigt dabei sowohl den konkreten Navigationsverlauf als auch das jeweilige Nutzerprofil: „Making dynamic recommendations to a Web user, based on her / his profile in addition to usage behavior is very attractive to many applications, e.g. cross-sales and up-sales in e-commerce" (Srivastava 2000: 6f.).

32 Vgl. Kap. 8, *Ereignisdatenanalyse*.

Dabei wird die zeitliche Abfolge von Elementen (z.B. Seitenaufrufen) analysiert. Die Analyse von Sequenzdaten mittels Optimal-Matching, wie sie in der vorliegenden Arbeit vorgeschlagen wird, bildet die Ausnahme. So verwenden Hay / Wets / Vanhoof (2001, 2002) das Verfahren des *Sequence Alignment* (SAM), um aufbauend auf einer Clusteranalyse eine Nutzertypologie zu entwickeln. Nutzerinnen werden darauf hin mit dem Ziel der Personalisierung des Angebotes den entwickelten Typologien zugeordnet. Hay et al. verwenden dabei jedoch einen spezifischen Algorithmus zur Sequenzanalyse, in den über die für den Optimal-Matching Algorithmus grundlegenden Operationen („insertion", „deletion", „substitution") hinaus die Operation „reorder" implementiert ist. Damit führt dieser spezielle Algorithmus zu Ergebnissen, die sich von denen der Optimal-Matching Analyse unterscheiden.[33]

Zusammenfassend kann festgehalten werden, dass ein gemeinsamer Ausgangspunkt des *Web-Mining* und der Navigationsanalyse in der *explorativen Analyse* der Navigation in Online-Umgebungen besteht mit dem Ziel des Identifizierens und Extrahierens von *Mustern, Regelmäßigkeiten und Strukturen*. Während jedoch beim *Web-Mining* ein sehr breites methodisches Spektrum verwendet wird, liegt die spezifische Fokussierung der Navigationsanalyse auf dem Navigationsverlauf als zusammenhängender Sequenz und nicht als zeitlicher Ablauf im Sinne des vorher – nachher einzelner Seiten (vgl. Kap. 5, *Sequenzdatenanalyse*). Methodisch bedeutet dies die Verwendung des Optimal-Matching Verfahrens zur Analyse des Navigationsverlaufs als Sequenz (vgl. 6, *Optimal-Matching Analyse*). Einen weiteren gemeinsamen Ausgangspunkt stellt die Verwendung von *Metadaten* zur Analyse von Navigationsverläufen dar. Jedoch unterscheidet sich das Vorgehen hinsichtlich wesentlicher Punkte: Bei dem Ansatz des *Web-Mining* und vor allem des *Web-Utilization Mining* werden besuchte Seiten als Elemente von Navigationssequenzen *im Nachhinein,* d.h. im Rahmen der Analyse auf der Grundlage von Ordnungssystemen klassifiziert. Bei dem Ansatz der vorliegenden Navigationsanalyse wird das Wissen *im Vorfeld* der Nutzung auf Grundlage einer *didaktischen Ontologie* (vgl. Kap. 3, *Web-Didaktik*) aufbereitet. Diese Aufbereitung des Wissens im Vorfeld hat für die Navigationsanalyse den entscheidenden Vorteil, dass die Informationen bereits in Hinblick auf ihre spätere Nutzung und Verwendung erzeugt werden und vermeidet damit insbesondere Probleme der eindeutigen Zuordnung im Rahmen nachträglicher Klassifikation. Dies ermöglicht eine eindeutige Interpretation der Inhalte auf der Ebene der Metadaten, im Gegensatz zu im Nachhinein erstellten und angewendeten Klassifikationssystemen.[34]

Über die Nutzungsdaten hinausgehenden Informationen (*user profile*s, *customer profile information)* stehen bei der vorliegenden Navigationsanalyse nicht zur Verfügung: Es ist jedoch grundsätzlich möglich die Ana-

33 Die detaillierte Beschreibung des Optimal-Matching Algorithmus folgt in Kapitel 6: 41.

34 Das Problem der Entwicklung handhabbarer Klassifikationsschemata für Web-Informationen zeigt sich deutlich im Bereich des *Web-Mining* als Disziplin des *Data-Mining* (vgl. Chakrabarti 2003) und der generellen Frage, wie die heterogenen Informationen des WWW indexiert werden können. Kennzeichnend ist dabei ein (automatisiertes) Indexieren von Webseiten im Nachhinein, das dem Vorgehen von Suchmaschinen entspricht.

lyse auf solche Daten auszuweiten, z.B. durch Online-Nutzerbefragungen nach Beendigung der Navigation in der Lernumgebung (z.B. im Rahmen einer Evaluation).

In Anlehnung an die oben genannten technischen Möglichkeiten der Rekonstruktion von Navigationsverläufen stellt sich die Frage, worin dieses Potential analog zum *Web-Mining* / *E-Commerce* für das *E-Learning* bestehen kann. Für den pädagogisch-didaktischen Bereich des E-Learning stehen Analysen von Navigationsverläufen sowie die Personalisierung von Lernangeboten und Lernumgebungen im Vergleich zum E-Commerce erst im Anfangsstadium. So gibt es beispielsweise kein pädagogisch-didaktisches Empfehlungssystem („recommendation system“) in Lernumgebungen, dass dem von *amazon.de* für den Bereich des E-Learning entspricht.[35] So ist davon auszugehen, dass ein Transfer von Empfehlungssystemen aus dem Bereich des E-Commerce in den Bereich des E-Learning konzeptionell-theoretisch sehr komplex ist. Die Aktivitäten eines *Käufers* unterscheiden sich von denen eines *Lernenden*, der *Kaufprozess* unterscheidet sich vom *Lernprozess*, das *Ergebnis* eines Lernprozesses unterscheidet sich vom *Ergebnis* einen Kaufprozesses. Darüber hinaus bleibt z.B. die Frage offen, ob ein Empfehlungssystem ähnliche Informationen vorschlägt (wie z.B. bei *amazon.de*) oder aber im Sinne einer absichtsvollen Irritation stark abweichende Informationen.

Auf die Unterschiede zwischen E-Commerce und E-Learning sowie auf das Fehlen entsprechender pädagogisch-didaktischer Empfehlungssysteme weist auch Zaïane (2001: 60) hin: „However, while there are clever tools developed to understand online customer's behavior in order to increase sales and profit, there is very little done to automatically discover access patterns to understand learners' behavior on webbased distance learning.“ In diesem Zusammenhang kritisiert Zaïane (2002) das eingeschränkte Potenzial der Entwicklung solcher Empfehlungssysteme auf der Grundlage der Analyse aggregierter Logdaten. Diese Daten haben nur sehr eingeschränkte Aussagekraft für das Verständnis von Navigationsprozessen und speziell von E-Learning-Prozessen. Daher arbeitet auch Zaïane (2001, 2002) mit dem Verfahren des Optimal-Matching, um das Navigationsverhalten im WWW zu analysieren.

Entsprechend den Ausführungen in diesem Abschnitt versteht sich die vorliegende Arbeit der explorativ-heuristischen Analyse empirischer Navigationsverläufe als eine der grundlegenden Methodologien und Ausgangspunkte der Entwicklung pädagogischer Empfehlungssysteme. Dabei ist in einem ersten Schritt grundsätzlich zu analysieren, wie die konkreten empirischen Navigationsverläufe aussehen, um daraufhin überhaupt erst Empfehlungen zu diesen Verläufen aussprechen zu können.

35 Unter einem Empfehlungssystem versteht Zaïane (2002: 59) ein Programm, das die Handlungen des Nutzers interpretiert und daraufhin Empfehlungen ausspricht, die für den Nutzer hilfreich sind: „A recommender system is a program that sees what a user is doing and tries to recommend courses of action it thinks would be beneficial to the user“. Für einen Überblick unterschiedlicher Empfehlungssysteme im Bereich E-Commerce, vgl. Srivastava (2000: 6f.).

5 Sequenzdatenanalyse

Im Kapitel 2 wurde die allgemeine methodologische Konzeption der *Navigationsanalyse* erläutert. Gegenstand dieses Kapitel ist die *Sequenzdatenanalyse* als ein Teilgebiet der *Navigationsanalyse*, das sich auf die quantitative Analyse von *Navigationsverläufen* (als spezifische Form von *Textdaten)* bezieht. In technischer Hinsicht basieren diese Sequenzdaten auf den serverseitig aufgezeichneten Logdaten der Nutzung von Online-Umgebungen. Aufbauend auf diesen Logdaten wird der Navigationsverlauf einzelner Nutzender als Sequenz rekonstruiert. Diese Sequenzen bilden dann den Ausgangspunkt der Sequenzdatenanalyse mittels Optimal-Matching. Das Ergebnis der Optimal-Matching Analyse ist eine Distanzmatrix, die wiederum den Ausgangspunkt für weitergehende strukturen-erklärende oder strukturen-entdeckende Verfahren bildet (vgl. Backhaus et al. 2000). Bei der vorliegenden Studie besteht diese weitergehende Analyse in der Clusterung von Navigationssequenzen auf Grundlage der Levenshtein-Distanzen, um ähnliche Sequenzen zu gruppieren. Die Navigationsanalyse basiert damit auf strukturen-entdeckenden Analyseverfahren.

Im Gegensatz zu statistisch-deskriptiven Verfahren der *Logdaten-Analyse* als Analyse aggregierter Daten (vgl. Kap.4.1: 27) und zu inferenzstatistischen und multivariaten Verfahren des *Web-Mining* (vgl. Kap. 4.2: 30) wird in diesem Kapitel das analytische Potenzial der Sequenzanalyse für die *Analyse vollständiger Sequenzen* mittels Optimal-Matching dargestellt. Einleitend wird der Begriff der „Sequenz“ näher erläutert. Daran anschließend wird die Verwendung der Methode der Sequenzdatenanalyse in den Sozialwissenschaften dargestellt.

Im nächsten Kapitel wird dann die Sequenzdatenanalyse auf Grundlage der Optimal-Matching Analyse dargestellt.

5.1 Der Begriff der „Sequenz“

Unter einer „Sequenz“ versteht man allgemein eine Abfolge, Reihung oder Reihenfolge von Elementen. Ein einzelnes, isoliertes Element stellt demnach keine Sequenz dar. Als Prototyp einer Sequenz in den *Naturwissenschaften* – insbesondere in der Molekularbiologie - gilt die DNA als Träger des menschlichen Erbgutes. Die Erbinformation ist in der *räumlichen Anordnung* von vier Aminosäuren als Grundelementen enthalten. Die Analyse und Entschlüsselung der DNA ist Gegenstand des *Human Genome Project* (HGP).[36]

36 vgl. Human Genome Project, <www.genome.gov>, (28.08.2006).

Als Prototyp einer Sequenz in den *Sozialwissenschaften* – insbesondere in der Soziologie - kann der Lebenslauf bezeichnet werden, wie er in der Lebensverlaufsforschung analysiert wird (vgl. Erzberger 2001; Erzberger / Prein 1997; Brüderl / Scherer 2005; Windzio 2001; Sackmann / Wingens 2001). Fokussiert wird dabei allgemein das „Zusammenspiel institutioneller und individueller Dynamiken über den Lebenslauf hinweg“ (Sackmann / Wingens 2001: 5), z.B. das Erwerbsleben von Personen, wobei der Erwerbsverlauf als *zeitlicher Verlauf* unterschiedlicher Zustände analysiert wird. Die folgende Abbildung 6 „Lebenslauf der 60er Kohorte“ veranschaulicht unterschiedliche Verläufe der Erwerbstätigkeit. In der ersten Spalte befindet sich die Nummerierung der einzelnen Fälle; in der zweiten Spalte die den Fällen entsprechenden Sequenzen.

Abb. 7: Lebensverläufe der 60er Kohorte (13 Statuspositionen)

Nr.	t_1 - t_{24}
1	ADDDDDDDDDDDFFFFF FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
2	BAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBB
3	
4	C-AAAAAIIIIIIIBBBICIICIIICIIIBBBBBBBBBBBBBBBBBBBBB
	JAAAAAIIBBBBFFFFF FFFFFFFFFFFBBBBBBBBBBBBBBFFAAAAII
.	.....
.	.....
530	

(Aufgeführt werden nur Positionen, die in diesen vier Verläufen auch vorhanden sind)

A Berufsarbeit (vollzeit)
B Berufsarbeit (teilzeit)
C Berufsarbeit unterhalb der Rentenversicherungspflicht
D berufsfremde Tätigkeiten (vollzeit)
E berufsfremde Tätigkeiten (teilzeit)
F berufsfremde Tätigkeiten unterhalb der Rentenversicherungspflicht
I Arbeitslosigkeit
J längere Krankheiten

Abbildung 6: „Lebenslauf der 60er Kohorte“, in: Erzberger (2001: 147)

Die Sequenz besteht in diesem Beispiel aus der Zuordnung von definierten Zuständen des Erwerbslebens zu definierten zeitlichen Intervallen:

- *Fall 1* befindet sich zu Beginn des Analysezeitraums für ein Zeitintervall im Zustand „Berufsarbeit (vollzeit)“. Danach wechselt Fall 1 für insgesamt 11 Zeitintervalle in den Zustand „berufsfremde Tätigkeiten (vollzeit)“, danach für insgesamt 39 Zeitintervalle in den Zustand „berufsfremde Tätigkeiten (teilzeit)“, etc.
- *Fall 2* befindet sich zu Beginn des Analysezeitraums für ein Zeitintervall im Zustand „Berufsarbeit (teilzeit)“. Danach wechselt Fall 2 für insgesamt 23 Zeitintervalle in den Zustand „Berufsarbeit (vollzeit)“, etc.
- Der Querstrich beim zweiten Zeitintervall des *Falls 3* drückt das Fehlen einer Zuordnung aus. Für dieses Zeitintervall ist keine Information über den Zustand vorhanden. Dieses Fehlen wird auch als *Lücke* („gap“) in der Sequenz (der Erwerbsarbeit) bezeichnet.

Innerhalb einer Sequenz wird der Wechsel zwischen zwei Zuständen (z.B. von „Berufsarbeit (vollzeit)“ zu „berufsfremde Tätigkeiten (vollzeit)“) *zu einem bestimmten Zeitpunkt t* als *Ereignis* (event) bezeichnet. Als *Übergang* wird der Wechsel zwischen zwei Zuständen in einem Prozess bezeichnet, der *mehr oder weniger Zeit* in Anspruch nehmen kann (vgl. Tabelle 8: 92).

Als *Verlauf* wird die Gesamtheit aller ermittelten Zustände inklusive der Verweildauer in den einzelnen Zuständen bezeichnet. Die Verweildauer in den einzelnen Zuständen wird in der obigen Abbildung 6 durch die wiederholte Zuordnung des betreffenden Zustandes zu folgenden Zeitintervallen ausgedrückt.

Als Prototyp einer Sequenz im Kontext der *Navigationsanalyse* wird der Navigationsverlauf eines Nutzers in einer Lernumgebung aufgefasst. Die Navigations*sequenz* als Gegenstand der Sequenzdatenanalyse besteht aus der *zeitlichen Abfolge besuchter Seiten*. Die besuchten Seiten bilden dabei die Elemente bzw. Zustände; der Zustandsraum allgemein besteht aus den Lerneinheiten bzw. den diesen zugeordneten Wissenseinheiten, die prinzipiell ausgewählt werden können.[37] Aus pädagogischer Perspektive kann der Navigationsprozess dabei als *selbstgesteuertes Lernen* oder genauer als „*autodidaktisches Handeln*“ (vgl. Meder 1995a, 1995b, 2006, s. auch Kap. 3.2, *Navigation als Autodidaktik*) interpretiert werden. Die kleinste mögliche Untersuchungseinheit bildet der Navigationsprozess innerhalb einer *Lerneinheit*, die auch als *Mikronavigation* bezeichnet wird. Dieser Prozess der Mikronavigation bildet auch den Gegenstand der in Kapitel 7 beispielhaft vorgestellten Sequenzanalyse.

5.2 Sequenzdatenanalyse in den Sozialwissenschaften

Im Bereich der *Naturwissenschaften* und dort vor allem im Bereich der *Molekularbiologie* stellt die Sequenzdatenanalyse seit der Analyse und Entschlüsselung der menschlichen DNA ein verbreitetes methodisches Vorgehen dar. Die Analyse so komplexer und langer Ketten von Elementen wie der menschlichen DNA wird erst ermöglicht durch den Einsatz von Computertechnologie und speziell durch den Einsatz von Algorithmen zur Analyse solcher Sequenzen.

Das Standard- und Referenzwerk der Sequenzdatenanalyse im naturwissenschaftlichen Bereich ist das von David Sankoff und Joseph Kruskal herausgegebene „Time Warps, String Edits, and Macromolecules“ aus dem Jahr 1983 (im Folgenden zitiert nach der 2. Auflage von 1999). Vor allem der Beitrag „An Overview of Sequence Comparison“ von Kruskal (1999: 1ff.) gibt einen zusammenfassenden Überblick über unterschiedliche Verfahren der Sequenzanalyse. Die zentrale Forschungsfrage formuliert Kruskal wie folgt: „Dealing with differences between sequences due to deletion – insertion [...] and substitution is the central theme of sequence comparison“ (Kruskal 1999: 9). Es geht also um die Beantwortung der Frage, wie auf Grundlage der Operationen *Löschen* („deletion“), *Einfügen* („insertion“) und *Austauschen* („substitution“) die Ähnlichkeit bzw. Unähnlichkeit von Sequenzen bestimmt werden kann. Den Kern der Sequenzdatenanalyse bildet

37 Der einzelne Aufruf lediglich einer Wissenseinheit stellt aus der Perspektive der Mikronavigation keine Sequenz dar. Bei einer erweiterten Perspektive auf Prozesse der Makronavigation als Navigation zwischen Lerneinheiten stellt jedoch auch dieser einzelne Aufruf ein Element einer übergeordneten Sequenz dar. Die Makroperspektive ist jedoch nicht Gegenstand der vorliegenden Arbeit.

der *paarweise Vergleich* aller Sequenzen mit dem Ziel, eine Ausgangssequenz mit Hilfe der grundlegenden Operationen in eine Zielsequenz zu transformieren, d.h. eine Übereinstimmung („alignment") der Ausgangs- und der Zielsequenz herzustellen.

Zum Forschungsfeld der Sequenzanalyse gehören einerseits die Entwicklung effizienter Methoden und Algorithmen zur Durchführung dieses paarweisen Sequenzvergleichs, sowie die Bestimmung der *geringsten Anzahl* der erforderlichen Operationen, um dieses „alignment" herzustellen („*optimal alignment*" bzw. „*optimum analysis*"). Diese geringste Anzahl der erforderlichen Operationen zu Herstellung der Übereinstimmung von Ausgangs- und Zielsequenz dient dann als *Maß der Distanz* zwischen Sequenzen, als Maßzahl für deren *Unähnlichkeit*.[38]

Als *Anwendungsgebiete* dieser zum Zeitpunkt der Veröffentlichung 1983 neuartigen technologischen Verfahren der Analyse von Sequenzen nennt Kruskal den Bereich

- der *Molekularbiologie*, und dort z.B. der Analysen der Homologie von Makromolekülen: „Are certain macromolecules homologous? Which parts of one molecule are homologous to which parts of the other? If we compare many different molecules in one species with the corresponding molecules in another species, how high is the typical degree of homology? For what pairs of species is this typical degree of homology high?" (Kruskal 1999: 3);
- der *Sprachforschung* („speech research"), und dort den Bereich der Sprecher- und der Spracherkennung: das Erkennen einzelner Worte, das Erkennen unbekannter Worte (die Bezeichnung „time-warping" aus dem Titel des Sammelbandes verweist auf das Feld der Sprachforschung und drückt die Notwendigkeit der Berücksichtigung des zeitlichen Faktors in Form von „compression" und „expansion" von Sequenzen bzw. der unterschiedlichen Aussprache und Betonung von Worten aus, vgl. Kruskal 1999: 4);
- der *Computerwissenschaften und Informatik*, und dort beispielsweise das als „string-correction" oder „string-editing" bezeichnete Problem des Vergleichs von Dateien bzw. deren Versionskontrolle;
- der *angewandten Chemie*, und dort beispielsweise die Gaschromatography, bei der es um physikalische Methoden der Trennung von Gasgemischen und deren Analyse geht;
- der *technischen Datenübertragung*, und dort das Feld der Codierung / Recodierung von Daten und der Fehlerkontrolle. Aus diesem Bereich stammen auch die grundlegenden Forschungen Levenshteins (1966).

Mathematische Grundlagen der Sequenzdatenanalyse werden im Rahmen des Sammelbandes von Sankoff / Kruskal besonders im Beitrag „On the Complexity of the Extended String-to-String Correction Problem"

38 Zur Erweiterung dieses Konzeptes der Anzahl der erforderlichen Operationen zur Gewichtung der Operationen durch *Kosten*, vgl. Kapitel 6.3: 48.

von Wagner (1983) diskutiert, sowie deren mathematische Bearbeitung durch dynamische, rekursive und iterative Prozesse und Programmiertechniken auf Grundlage unterschiedlicher Algorithmen.

Die Verwendung von Methoden der Sequenzdatenanalyse stellt für den Bereich der *Sozialwissenschaften* ein junges methodisches Vorgehen dar. Der Transfer dieser Methode aus dem Bereich der Naturwissenschaften speziell in den Bereich der Soziologie geht auf Andrew Abbott zurück, dessen Forschungen seit Ende der 1980er Jahre in methodischer Hinsicht „Pioniercharakter" (Aisenbrey 2001: 43, vgl. auch Halpin / Chan 1998) bezüglich der Analyse sozialwissenschaftlicher Sequenzdaten zukommt.

Ausgangspunkt der Diskussion im sozialwissenschaftlichen Bereich bildet die Veröffentlichung der Studie „Optimal Matching Methods for Historical Sequences" von Abbott / Forrest (1986), in der das Potenzial der Sequenzdatenanalyse auf Grundlage des Verfahrens des Optimal-Matching am Beispiel der Entwicklung und des Vergleichs von Tanzschritten demonstriert wurde. Diese Studie bildet gleichzeitig den Ausgangspunkt der kritischen Diskussion, die sich exemplarisch in der „Special Section on Sequence Analysis" der Zeitschrift *Sociological Methods and Research* (2000) rekonstruieren lässt (vgl. Abbott2000a; Abbott 2000b; Levine 2000; Wu 2000; Dijkstra / Toon 1995; Toon 2000; Driel / Oosterveld 2001; Elzinga 2003). Vor allem seit den 1990er Jahren veröffentlicht Abbott mit unterschiedlichen Kollegen Studien auf Grundlage der Optimal-Matching Methode und gilt als der prominenteste Vertreter dieses Ansatzes in den Sozialwissenschaften.

Für einen allgemeinen Überblick über die Forschungsaktivitäten der Veröffentlichungen Abbotts et al. zur Optimal-Matching Analyse wird an dieser Stelle auf das Literaturverzeichnis verwiesen (Abbott/ Forrest 1986, Abbott 1990a, Abbott 1990b; Abbott 1990c; Abbott / Hrycak 1990, Abbott1992, Abbott 1995a, Abbott / Barman 1997, Abbott / Tsay 2000), sowie auf Aisenbrey (2000) und Brüderl / Scherer (2005) für einen allgemeinen Überblick über den Einsatz im sozialwissenschaftlichen Bereich.

Die folgende kursorische Auswahl von Forschungsarbeiten auf Grundlage der Sequenzanalyse verdeutlicht das vielfältige Themenspektrum des Einsatzes der Optimal-Matching Methode:

- „Optimal Matching Methods for Historical Sequences" (Abbott / Forrest 1986);
- „Measuring Resemblances in Sequence Data: An Optimal Matching Analysis of Musicians' Careers" (Abbot / Hrycak 1990);
- „Local Sequential Patterns: The Structure of Lynching in the Deep South, 1882 - 1930" (Stovel 2001);
- „Optimal Matching Analysis: A methodological note on studying career mobility" (Chan 1995);
- „Ascription into Achievement. Models of Career Systems at Lloyds Bank, 1890-1970" (Stovel / Savage / Bearman 1996) ;
- „Optimal-Matching-Technik: Ein Analyseverfahren zur Vergleichbarkeit und Ordnung individuell differenter Lebensverläufe" (Erzberger / Prein 1997);

- „Class Careers as Sequences: An Optimal Matching Analysis of Work-Life-Histories“ (Halpin / Chan 1998);
- „Clocking Out: Multiplex Time in Retirement“ (Han / Moen 1998);
- „Early Career Patterns: A Comparison of Great Britain and West Germany“ (Scherer 1999);
- „Übergangsmuster in der Statuspassage von beruflicher Bildung in den Erwerbsverlauf“ (Mowitz-Lambert 2001);
- „Übergänge und Sequenzen. Der Einfluss der Arbeitslosigkeit auf den weiteren Erwerbsverlauf“ (Windzio 2001);
- „Residential Trajectoires: Using Optimal Alignment to reveal the Structure of Residential Mobility“ (Stovel / Bolan 2004);
- „Die Pluralisierung partnerschaftlicher Lebensformen in Westdeutschland“ (Brüderl 2004).

Zusammenfassend ist festzuhalten, dass diese Forschungsarbeiten überwiegend aus dem Bereich der Soziologie stammen, und genauer: der soziologischen Lebensverlaufsforschung. In der pädagogischen Forschung findet die Methode der Sequenzanalyse mittels Optimal-Matching, so wie sie in dieser Arbeit ausgeführt wird, bisher keine Verwendung.[39] Auch für den Bereich der Analyse von Navigationsverläufen in Online-Lernumgebungen unter pädagogisch-didaktischer Perspektive liegen bisher keine Untersuchungen auf dieser methodologischen Grundlage vor.

Das methodische Vorgehen der Seqenzdatenanalyse kann bei allen Unterschieden der konkreten Durchführung im Detail in einer allgemeinen Perspektive wie folgt skizziert werden:

- *Konzeption*: Definition des Zeitraumes (Beobachtungsfenster), der Zeitachse der Datenerhebung (Erhebungsabstände), des Erhebungsgegenstandes; Konstruktion des Zustandsraumes.
- *Erhebung von Sequenzdaten* (retrospektive Konstruktion von Sequenzen aus vorliegenden Datenquellen bzw. Erhebung prozessgenerierter Daten, vgl. Baur 2005).
- *Kodierung* der erhobenen Daten in Form von Zeichen-Sequenzen, anhand eines definierten Zeichen-Alphabets;
- *Sequenzdatenanalyse mittels Optimal-Matching* (vgl. Kap.6; *Optimal-Matching Analyse*): die einzelnen Sequenzen werden paarweise verglichen. Ergebnis dieses Vergleichs ist die *Maßzahl* der Levenshtein-Distanz, in der die Distanz der verglichenen Sequenzen ausgedrückt wird. Die Levenshtein-Distanz als Ergebnis des paarweisen Sequenzverleichs wird in Form einer *Distanzmatrix* abgebildet und stellt den Ausgangspunkt für weitere konfirmatorische (strukturen-prüfende) oder heuris-

39 Ein Beispiel des Einsatzes der Analyse von Sequenzen im Bereich der Pädagogik stellt die Untersuchung temporaler Muster im Bereich der Freizeitforschung dar (vgl. Dollase u.a. 2000). Jedoch beruht das methodische Vorgehen vorrangig auf der Ereignisdatenanalyse (vgl. Kap. 8) und nicht auf der Sequenzanalyse. Das verwendete Verfahren des „Sequence Alignment“ (SAM) arbeitet nicht mit den Operationen INDEL und Substitution, sondern mit dem Einfügen von Leerstellen („gaps“), um eine Übereinstimmung der Ausgangs- und Zielsequenz zu erreichen.

tisch-explorative (strukturen-entdeckende) Analysen dar (vgl. Kap. 6.7; *Potential der Optimal-Matching Analyse*).

Im Folgenden wird die Konzeption und Durchführung der Sequenzdatenanalyse als heuristisch-explorative Analyse mittels Optimal-Matching dargestellt. Die konfirmatorische Analyse von Sequenzdaten ist nicht Gegenstand der vorliegenden Arbeit.

6 Optimal-Matching Analyse

Der Begriff der *Optimal-Matching Analyse* (OMA) wird in dieser Arbeit für Verfahren verwendet, die auf Grundlage der Levenshtein-Distanz und der Operationen *Substitution* und *Indel* („insertion" und „deletion") sowie der Gewichtung von Operationen durch *Kosten* unter Verwendung iterativer Prozeduren (Algorithmen) die Distanz von Sequenzen bestimmen (vgl. Elzinga 2005: 3).

Allgemeines Ziel der Sequenzdatenanalyse ist der Vergleich von Sequenzen, um *Muster, Regelmäßigkeiten und Strukturen* zu erkennen. Dieses Ziel kann auf unterschiedliche Art und Weise umgesetzt werden, mit Hilfe unterschiedlicher Algorithmen und auf Grundlage unterschiedlicher Distanz-Konzepte (vgl. Kruskal 1999). In diesem Kapitel wird *Optimal-Matching* (OM) als *ein spezifisches* methodisches Vorgehen der Analyse von Sequenzdaten ausgeführt, als *eine spezifische Verfahrensweise (*Algorithmus).

Eine softwaretechnische Umsetzung findet der OM-Algorithmus in dem Programm „Transition Data Analysis" (TDA), einem Programm zur statistischen Datenanalyse, das an der Fakultät für Sozialwissenschaften der Ruhr-Universität Bochum von Götz Rohwer und Ulrich Pötter entwickelt wird.[40] Der Schwerpunkt von TDA liegt auf der Analyse von Übergängen, die in Kapitel 8 als Ereignisdatenanalyse beschrieben wird. Darüber hinaus ist in TDA auch der Optimal-Matching Algorithmus zur Berechnung der Levenshtein-Distanz implementiert.[41]

In der Literatur wird oft synonym zum Begriff der *Optimal-Matching Analyse* (OMA) der Begriff *Sequence Alignment Method* (SAM) verwendet. Diese synonyme Verwendungsweise kritisiert Elzinga (2003: 4), da unklar bleibt, ob der spezifische Algorithmus des *Optimal-Matching* gemeint ist oder aber das methodologische Konzept der Sequenzanalyse als solches. In dieser Arbeit bezeichnet „Sequenzdatenanalyse" die übergeordnete Methodologie und „Optimal-Matching" einen konkreten Algorithmus zur deren Umsetzung.[42]

Auf welche Weise werden nun auf Grundlage des Optimal-Matching Sequenzen verglichen? Wie wird deren Ähnlichkeit bzw. Unähnlichkeit festgestellt? Im Folgenden wird das Optimal-Matching Verfahren zur Analyse von Sequenzdaten näher dargestellt.

Einleitend werden mit der Hamming- und der Levenshtein-Distanz zwei unterschiedliche Verfahren zur Bestimmung der Distanz von Sequenzen gegenüber gestellt (Kap. 6.1).

40 TDA steht als Freeware unter den Bedingungen der *GNU General Public Licence* (GPL) zur Verfügung, <http://www.stat.ruhr-uni-bochum.de/tda.html>, (28.08.2006).

41 Das von Andrew Abbott entwickelte Programm „OPTIMIZE" zur Sequenzanalyse mittels Optimal-Matching wird in der Navigationsanalyse nicht verwendet, da die Analyse einer so großen Anzahl von Navigationssequenzen wie im vorliegenden Fall mit dieser Software nicht möglich ist.

42 Vor allem im Bereich der Naturwissenschaften findet sich eine Vielzahl unterschiedlicher Algorithmen zu Analyse von Sequenzdaten (vgl. Sankoff / Kruskal 1999).

Daran anschließend wird die Berechnung der Levenshtein-Distanz als zentrales Element der Optimal-Matching Analyse erläutert (Kap. 6.2).

Anhand der Gewichtung von Operationen durch *Kosten* wird eine Weiterentwicklung und Spezifizierung der Optimal-Matching Analyse vorgestellt (Kap. 6.3), gefolgt von allgemeinen Überlegungen zur Definition von Substitutions- bzw. Indelkosten (Kap. 6.4) und deren Relation (Kap. 6.5) .

Abschließend werden Möglichkeiten des Umgangs mit unterschiedlich langen Sequenzen erläutert (Kap. 6.6), sowie das grundsätzliche Potential der Optimal-Matching Analyse (Kap. 6.7).

6.1 Distanzmaße: Hamming und Levenshtein

In diesem Abschnitt wird am Beispiel der Hamming- und der Levenshtein-Distanz der Frage nachgegangen, auf welche Weise Sequenzen grundsätzlich miteinander verglichen werden können. Wie kann festgestellt werden, ob sich zwei Sequenzen ähneln oder unterscheiden? Wie kann festgestellt werden, wie stark sich Sequenzen ähneln oder unterscheiden?

Das Konzept der Hamming-Distanz (Hamming 1950) stammt aus dem Bereich der elektronischen Datenübertragung und stellt dort eine der bekanntesten Methoden zur Analyse von Sequenzen dar. Sie beruht auf einem direkten Vergleich der Anzahl von unterschiedlich besetzten Positionen zweier Sequenzen: es wird bei diesem Verfahren schrittweise verglichen, ob sich *gleiche Elemente* in sich *entsprechenden Positionen* befinden. Die Distanz zweier Sequenzen besteht dann in der Anzahl der mit unterschiedlichen Elementen besetzten Positionen (vgl. Kruskal 1999: 1f.). Unberücksichtigt bleiben bei diesem Vergleich gemeinsame Reihenfolgen oder Muster innerhalb der Sequenzen.

Allerdings kann in vielen sozialwissenschaftlichen Anwendungsgebieten – wie auch bei der Analyse von Navigationsprozessen – von einer solchen *inhärenten Korrespondenz* zwischen Sequenzen nicht ausgegangen werden. Ihr Vorhandensein ist in den meisten Fällen zunächst auch unklar, weil entweder zwischen den Sequenzen kein expliziter Zusammenhang besteht, oder weil die Sequenzen unterschiedliche Längen besitzen. Für diese Anwendungsgebiete sind daher differenziertere Verfahren des Vergleichs von Sequenzen erforderlich, die eine Analyse von Sequenzen ungleicher Länge und ohne inhärenter Korrespondenz ermöglichen.

Gerade die Berücksichtigung von Mustern, Regelmäßigkeiten und Strukturen bildet den Ausgangspunkt der Überlegungen von Levenshtein (1966) zum Vergleich von Sequenzen. Die Distanz von Sequenzen als Grad der Unähnlichkeit wird dabei errechnet aus der Anzahl der Transformationsschritte[43] die notwendig sind, um

43 In der vorliegenden Arbeit werden die Begriffe „Transformation“ und „Operation“ synonym verwendet. Diese (Bearbeitungs)Operationen werden auch als „Edit-Operation“ bezeichnet. Die Levenshtein-Distanz wird auch synonym als „Edit-

eine Ausgangssequenz in eine Zielsequenz zu überführen. Ziel dieser Transformationsschritte ist ein *alignment* (Abgleich, gleiche Anordnung).[44]

Dabei verwendet Levenshtein zwei Varianten des Distanz-Konzeptes:

- Das *erste Distanz-Konzept* errechnet die Distanz auf Grundlage der Anzahl der Transformationen *Einfügen* („insertion"), *Löschen* ("deletion") und *Austauschen* ("substitution") von Elementen.
- Das *zweite Distanz-Konzept* errechnet die Distanz allein auf Grundlage der Transformationen *Einfügen* („insertion") und *Löschen* ("deletion") von Elementen; das *Austauschen* ("substitution") stellt keine zulässige Operation dar.

Im Folgenden werden die Operation „insertion" und „deletion" in Anlehnung an die allgemeine Fachterminologie als „Indel" (*in*sertion *del*etion) zusammengefasst. Diese Zusammenfassung beruht auf der Komplementarität dieser beiden Operationen: das Einfügen eines Elementes in einer Ausgangssequenz ist gleichbedeutend mit dem Löschen eines Elementes in eoner Zielsequenz. Generell gilt, dass sich die verglichenen Sequenzen umso ähnlicher sind, je weniger Transformationsschritte benötigt werden, um ein „alignment" der Sequenzen zu erreichen.

In der Verwendung der grundlegenden Transformationen *Einfügen*, *Löschen* und *Austauschen* für die Analyse von Sequenzen kommt eine spezifische Konzeption von „Distanz" zwischen Sequenzen zum Ausdruck, die wesentlich differenzierter als die Hamming-Distanz ist, da grundsätzlich berücksichtigt wird, ob die zu vergleichenden Sequenzen gemeinsame Muster, Regelmäßigkeiten und Strukturen enthalten.

Aus Sicht des *Optimal-Matching* handelt es sich bei der oben beschriebenen Hamming-Distanz um einen Spezialfall des ersten Distanz-Konzeptes als einem *Substitutionsalgorithmus*, bei dem das *Austauschen* ("substitution") von Elementen die einzig zur Verfügung stehende Operation darstellt.[45]

In Hinblick auf das zweite Distanz-Konzept handelt es sich bei der Hamming-Distanz um ein entgegengesetztes Vorgehen: das *Austauschen* ("substitution") stellt die einzig zulässige Operation dar; das *Einfügen* („insertion") oder das *Löschen* ("deletion") von Elementen ist nicht zulässig.

Analog zum Gegenstand der Sequenzdatenanalyse (vgl. Kap. 5.2: 37) besteht der Optimal-Matching Algorithmus aus zwei *Prozessen:* Aus der allgemeinen Bestimmung aller Transformationsoperationen, um eine Quellsequenz in die Zielsequenz zu überführen („alignment"); sowie der Ermittlung der geringsten Anzahl der dazu notwendigen Operationen, d.h. der Analyse der minimalen Distanz zwischen Sequenzen durch einen iterativen Prozess der Minimierung („optimum analysis").

Distance" bezeichnet.

44 Daher wird in der Literatur die Methode des „Optimal-Matching" oft in einem synonymen Sprachgebrauch auch als „Sequence Aligment Method" (SAM) bezeichnet.

45 Das Ergebnis des Algorithmus von Hamming wird von Brüderl / Scherer (2005: 4) auch als „naive Distanz" bezeichnet.

6.2 Berechnung der Levenshtein-Distanz

Die *Vorgehensweise* des Optimal-Matching Algorithmus zur Bestimmung der Levenshtein-Distanz zwischen zwei Sequenzen wird in diesem Abschnitt anhand einer Matrix dargestellt. Das folgende Beispiel und die Grafiken stammen aus Erzberger (2001); weitere Bespiele finden sich in Abbott / Hrycak (1990), Erzberger / Prein (1997), Aisenbrey (2000); Brüderl / Scherer (2005), Elzinga (2003) und Elzinga (2005). Eine ausführliche Darstellung der *mathematischen Grundlagen* des Optimal-Matching Algorithmus findet sich bei Levenshtein (1966), Kruskal (1993), Wagner (1999); Sankoff / Kruskal (1999), Gusfield (1999); Clote / Backofen (2000).

Ausgangspunkt der folgenden Ausführungen bildet der Vergleich der Sequenzen a=(A, B, C, D, E, F) mit der Sequenz b=(B, C, D, E, F, A).

Die Unzulänglichkeit eines Substitutionsalgorithmus wie der oben beschriebenen Hamming-Distanz wird an diesem Beispiel offensichtlich: Es sind *keine gleichen Elemente in sich entsprechenden Positionen* vorhanden. Daher stellt der Hamming-Algorithmus für den Vergleich dieser Sequenzen eine maximale Ungleichheit (maximale Distanz) fest. Jede Position in Sequenz a (Quellverlauf) muss durch das entsprechende Element der Sequenz b (Zielverlauf) substituiert werden. Um eine Übereinstimmung herbeizuführen sind also 6 Operationen (in der Terminologie Erzbergers „Aktionen“, s. Abbildung 7: 44) erforderlich.

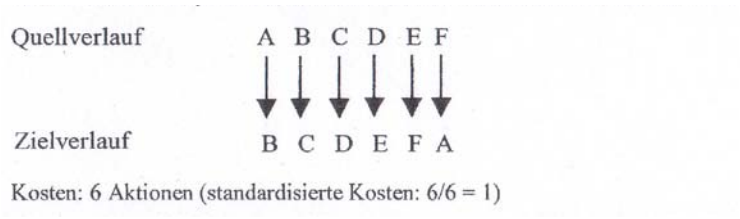

Abbildung 7: Überführung eines Quellverlaufs in einen Zielverlauf durch Ersetzen, in: Erzberger (2001: 147).

In diesen einfachen und überschaubaren Sequenzen ist jedoch eine gemeinsame Regelmäßigkeit bzw. ein gemeinsames Muster erkennbar: Quell- und Zielsequenz sind gegeneinander verschoben. Diese Struktur ist für einen Substitutionsalgorithmus jedoch nicht identifizierbar, sondern wird erst mit der Verwendung weiterer Operationen wie Einfügen („insertion“) und Löschen („deletion“) berücksichtigt.

Die geringste Anzahl von Operationen zum „alignment“ der beiden Sequenzen ergibt sich, wenn zunächst im Zielverlauf an der ersten Position ein A eingefügt („insertion“) wird (vgl. Abbildung 8: 45). Damit *verschieben* sich alle folgenden Positionen, woraus eine Übereinstimmung der Elemente B, C, D, E und F erfolgt. Abschießend muss lediglich das letzte Element des Zielverlaufs (A) gelöscht werden („deletion“), um eine Übereinstimmung des Quellverlaufs mit dem Zielverlauf herzustellen.

Die Unähnlichkeit zweier Sequenzen wird in Form einer Maßzahl ausgedrückt, der Levenshtein-Distanz. Damit wird genau die Distanz bezeichnet, die auf der *geringsten Anzahl* von Operationen beruht, um ein „alignment“ zu erreichen. Mit Hilfe des Optimal-Matching Algorithmus wird daher aus allen potenziell möglichen Transformationen, die zu einem „alignment“ der beiden Sequenzen führen, genau die Transformation ermittelt, die aus der geringsten Anzahl von Transformationsschritten besteht.

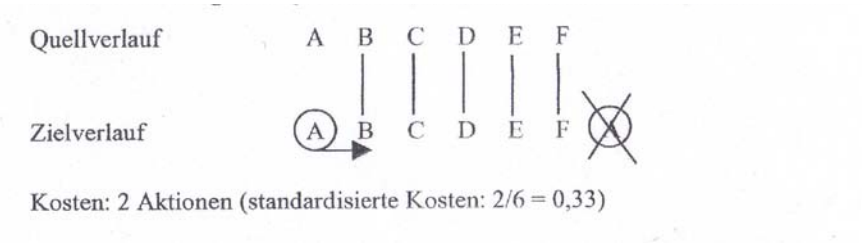

Abbildung 8: Überführung eines Quellverlaufs in einen Zielverlauf durch Einfügen / Löschen, in: Erzberger (2001: 148).

Im Gegensatz zur Hamming-Distanz mit 6 Operationen und dem Ergebnis der maximalen Distanz beider Sequenzen sind auf Grundlage des Optimal-Matching Algorithmus lediglich 2 Operationen notwendig. Die Sequenzen sind demnach nicht maximal verschieden, wie das Ergebnis der Hamming-Distanz nahe legt, sondern weisen eine ähnliche Struktur auf, was durch die Maßzahl der Levenshtein-Distanz zum Ausdruck kommt. Im Gegensatz zum Hamming-Algorithmus ist also der Optimal-Matching Algorithmus aufgrund der Operationen *Substitution* sowie *Indel* in der Lage, Regelmäßigkeiten innerhalb der zu vergleichenden Sequenzen zu identifizieren.

Das konkrete Vergehen der Ermittlung der Levenshtein-Distanz kann anhand einer Matrix verdeutlicht werden.[46] *Ausgangspunkt* der Transformationen ist die Zelle links oben, Zielpunkt die Zelle rechts unten. Der Algorithmus *spielt* nun unterschiedliche Möglichkeiten durch, um in der Matrix von links oben nach rechts unten zu gelangen und damit die Übereinstimmung der Elemente der Sequenzen zu erreichen.

Generell gibt es in jeder Zelle drei Möglichkeiten, zur nächsten Zelle zu gelangen.

- *Nach rechts*: diese Transformation entspricht dem *Einfügen* („insertion“) eines Elementes.
- *Nach unten*: diese Transformation entspricht dem *Löschen* („deletion“) eines Elementes.
- *Diagonal nach rechts unten*: diese Transformation entspricht dem *Austauschen* („substitution“) eines Elementes.

In der Tabelle ist mit Pfeilen der Weg eingezeichnet, der die Ermittlung der Levenshtein-Distanz als geringste Anzahl von Transformationen darstellt. Dabei kann es unterschiedliche Wege geben, um ein „align-

46 Grundsätzlich werden im Rahmen der Optimal-Matching unterschiedliche Matrizen verwendet: Eine Matrix zur Veranschaulichung der Ermittlung der Levenshtein-Distanz (vgl. Abbildung 9: 46); eine Matrix, die das Ergebnis des paarweisen Vergleichs aller Sequenzen enthält (vgl. Tabelle 4: 73, sowie 6: 81), sowie eine Substitutionskostenmatrix (vgl. Tabelle 22: 52), in der die unterschiedlichen *Kosten* für die Substitution von Elementen detailliert aufgeführt.

ment“ zu erreichen, für die jedoch die gleiche Anzahl von Transformationsschritten aufgewendet werden müssen. Für die Bestimmung der Maßzahl der Levenshtein-Distanz ist das Vorhandensein von mehr als einem Weg mit gleicher Anzahl von Transformationen unerheblich.

In einem *ersten Schritt* wird bei der Errechnung der Levenshtein-Distanz die Quellsequenz (Quellverlauf) in die erste Zeile der Matrix eingetragen; die Zielsequenz (Zielverlauf) in die erste Spalte.

	Quell-verlauf	A	B	C	D	E	F
Ziel-	Start	- -	- -	- -	- -	- -	- -
verlauf	- 0	1 1	1 2	1 3	1 4	1 5	1 6
B	- -	1 1	0 1	1 1	1 1	1 1	1 1
	- 1	1 1	1 1	1 2	1 3	1 4	1 5
C	- -	1 1	1 1	0 1	1 1	1 1	1 1
	2	1 2	1 2	1 1	1 3	1 4	1 5
D	- -	1 1	1 1	1 1	0 1	1 1	1 1
	- 3	1 3	1 3	1 2	1 1	1 2	1 3
E	- -	1 1	1 1	1 1	1 1	0 1	1 1
	- 4	1 4	1 4	1 3	1 2	1 1	1 2
F	- -	1 1	1 5	1 1	1 1	1 1	0 1
	- 5	1 5	1 1	1 4	1 3	1 2	1 1
A	- -	0 1	1 1	1 1	1 1	1 1	1 1
	- 6	1 5	1 6	1 5	1 4	1 3	1 2

Zellen in der Tabelle:

Kosten für Ersetzen ↘	Kosten für Löschen ↓
Kosten für Einfügen →	minimale Gesamtkosten

Abbildung 9: Matrix zur Ermittlung der Levenshtein-Distanz (Erzberger 2001: 149)

In einem *zweiten Schritt* werden in die einzelnen Zellen der Tabelle die *Kosten*[47] für die grundlegenden Operationen eingetragen (unter Verwendung des ersten Distanz-Konzeptes von Levenshtein werden in Abbildung 9: 46 alle Operationen einheitlich gewichtet, d.h. mit dem Wert „1“ versehen:[48] Die Levenshtein-Distanz errechnet sich dann aus der *Anzahl* der verwendeten Operationen.

47 Zur Erweiterung des Distanz-Konzeptes von Levenshtein durch die Definition von *Kosten* als Form der Gewichtung von Operationen, vgl. Kap. 6.3: 48.

48 In der Matrix befindet sich ein Tippfehler: Im Schnittpunkt der Spalte „B“ und der Zeile „F“ ist in der rechten oberen Ecke der Zelle als Kosten für die Operation *Einfügen* der Wert „5“ eingetragen. Wie in den anderen Zellen erkennbar beträgt dieser Wert „1“.

Entsprechend der oben beschriebenen Möglichkeiten von einer Zelle zur nächsten zu gelangen, stehen die

- Kosten für das *Austauschen* („substitution“) eines Elementes in der linken oberen Ecke der jeweiligen Zelle;
- die Kosten für das *Löschen* („deletion“) eines Elementes in der rechten oberen Ecke der jeweiligen Zelle;
- die Kosten für das *Einfügen* („insertion“) eines Elementes in der linken unteren Ecke der jeweiligen Zelle.

In einem *dritten Schritt* wird in die rechte untere Ecke der jeweiligen Zelle der Wert der *geringsten Kosten* eingetragen, die zum Erreichen dieser Zelle aufgebracht werden müssen.

Besondere Beachtung verdienen die Zellen der Tabelle, in deren linker oberer Ecke der *Wert „0“* steht. Dies bedeutet, dass für ein „alignment“ dieser Zellen keine Operationen notwendig sind, da die Elemente auf der y-Achse (Zielverlauf) und x-Achse (Quellverlauf) identisch sind (B-B, C-C, D-D, E-E, F-F).

An den eingezeichneten Pfeilen als konkretem Weg zur Bestimmung der Levenshtein-Distanz wird darüber hinaus deutlich, auf welche Weise der OM-Algorithmus Muster in den Sequenzen identifizieren kann: Indem das erste Element des Quellverlaufs gelöscht wird (dies entspricht einer ersten Operation) kann in den folgenden Schritten dem in beiden Sequenzen vorhandenen Muster gefolgt werden (da sich die Elemente des Musters im Quell- und Zielverlauf entsprechen sind dafür keine Transformationsoperationen notwendig) und es wird schließlich als letzter Schritt lediglich das „A“ als letztes Element gelöscht (dies entspricht einer weiteren Operation). Unter Ausnutzung des in beiden Sequenzen enthaltenen Musters sind im Rahmen des OM-Algorithmus lediglich zwei Operationen notwendig, um ein „alignment“ der beiden zu vergleichenden Sequenzen herzustellen.

Ausschlaggebend für die Levenshtein-Distanz als *geringste Anzahl* der Operationen der Transformation einer Quellsequenz in eine Zielsequenz ist die rechte untere Zelle der Tabelle: Dabei zeigt die Zahl, die innerhalb dieser Zelle an der rechten unteren Ecke steht, die *Levenshtein-Distanz* der zwei verglichenen Sequenzen an (in der Matrix in Abbildung 9: 46 der mit einem Kreis markierte Wert „2“).

Grundsätzlich kann die Maßzahl der Levenshtein-Distanz als *absolutes* oder als *relatives Distanzmaß* angegeben werden: Das *absolute Distanzmaß* bezieht sich auf die *geringste Anzahl* der verwendeten Operationen zum Erreichen einer Übereinstimmung („alignment“). Da die Anzahl der erforderlichen Operationen oft nicht unabhängig von der Gesamtlänge der Sequenzen interpretiert werden kann, wird aus Gründen der Vergleichbarkeit ein *relatives Distanzmaß*[49] verwendet. Dieses wird errechnet, indem die geringsten Kosten durch die Anzahl der Elemente der längeren Sequenz des paarweisen Sequenzvergleichs dividiert wird (vgl. Abbott / Hrycak 1990; Stovel et al. 1996). Der Wert dieses *relativen Distanzmaßes* liegt dabei zwischen „0“ und „1“, wobei „0“ vollkommene Gleichheit der Sequenzen und „1“ vollkommene Ungleichheit (d.h. maxi-

49 Erzberger (2001) spricht in diesem Zusammenhang von „standardisierten Kosten“.

male Distanz) bedeutet. Für die oben beschrieben Matrix ergibt sich ein *absolutes* Distanzmaß von „2“ bzw. ein *relatives* Distanzmaß von „0,33“ (2 Operationen und 6 Elemente der Sequenz; 2 / 6).

Im Rahmen der Optimal-Matching Analyse wird jede Sequenz des Datensatzes mit jeder anderen Sequenz verglichen (paarweiser Vergleich). Die Optimal-Matching Analyse bzw. der Optimal-Matching Algorithmus beantwortet Fragen nach Mustern, Regelmäßigkeiten und Strukturen von Sequenzen nicht auf direkte Weise, sondern liefert eine Maßzahl der Distanz für jeden paarweisen Sequenzvergleich. Diese Levenshtein-Distanzmatrix bildet dann den Ausgangspunkt für weitere Analysen (vgl. Kap. 6.7; *Potential der Optimal-Matching Analyse*).

Die errechneten Maßzahlen der Levenshtein-Distanz für jeden paarweisen Vergleich werden in eine Matrix, die sogenannte Levenshtein-Distanzmatrix eingetragen.[50] In der ersten Spalte sowie der ersten Zeile dieser Matrix befinden sich die Nummerierungen der zu vergleichenden Sequenzen. Die Diagonale von links oben nach rechts unten enthält dabei keine Distanzwerte, da die jeweilige Sequenz nicht mit sich selbst verglichen wird. Diese Diagonale stellt gleichzeitig eine Symmetrie-Achse dar, denn die Distanz von Sequenz a zu Sequenz b entspricht der Distanz von Sequenz b zu Sequenz a, diese spiegeln sich also entlang der Diagonalen. Diese symmetrische Relation ist ja gerade gemeint, wenn man von Ähnlichkeit spricht: a ist zu b genau so ähnlich wie b zu a (vgl. Elzinga 2003: 8). Diese Symmetrie kommt auch in der mathematischen Formel zur Berechnung der Anzahl der paarweisen Sequenzvergleiche zum Ausdruck: n * (n-1) / 2. Die Levenshtein-Distanz der Sequenz a – b ist also identisch mit der Sequenz b – a.

Ein Vergleich von 100 Sequenzen beruht im Rahmen der Optimal-Matching also auf 4950 paarweisen Sequenzvergleichen (100 * 99 / 2 = 9900 / 2 = 4950), die in die Levenshtein-Distanzmatrix eingetragen werden. Diese Beispielrechnung verdeutlicht gleichzeitig den hohen Rechenaufwand des Optimal-Matching Verfahrens, gerade bei einer großen Anzahl sowie sehr langer Sequenzen.

6.3 Gewichtung von Operationen durch *Kosten*

Auf die grundlegende Bedeutung des Konzeptes der Levenshtein-Distanz für das gesamte Feld der Sequenzdatenanalyse weist bereits Kruskal (1999: 5) hin: „His [Levenshteins, S.I.] distance function and generalizations of it play a major role in sequence comparison“. Nachdem im vorangehenden Abschnitt die Konzeption und Arbeitsweise des Optimal-Matching Algorithmus zur Berechnung der Levenshtein-Distanz dargestellt wurde, wird in diesem Abschnitt in Anlehnung an die von Kruskal erwähnten „generalizations“ die wichtigste Erweiterung dieses grundlegenden Konzeptes dargestellt: die Gewichtung von Operationen durch die Definition von *Kosten*.

50 Ein Beispiel für eine Levenshtein-Distanzmatrix befindet sich in Kapitel 7.1: 73.

Während Levenshteins ursprüngliche Konzeptionen von Distanz auf der *Anzahl* von Operationen beruht, wird mit der Erweiterung durch die Definition von *Kosten* eine Gewichtung von Operationen möglich (vgl. Elzinga 2005: 3f.). Für die oben erläuterte Levenshtein-Distanz bedeutet diese Erweiterung, dass nicht die minimale *Anzahl* der Operationen entscheidend ist, sondern die minimalen *Kosten* der erforderlichen Operationen. Das Konzept der Minimierung der Anzahl der Operationen wird also erweitert zur Maximierung der Kosteneffizienz; das Konzept des kürzesten Weges wird erweitert zu dem des kostengünstigsten Weges.

Hauptargument für die Verwendung der Gewichtung von Operationen durch *Kosten* ist dabei die auf eine inhaltliche Interpretation von Sequenzen bezogene Überlegung, dass nicht allen Operationen die gleiche Bedeutung zukommt, also inhaltlich nicht als gleich zu bewerten sind. In der Definition von *Kosten* kommen somit im Vorfeld der Analyse getroffene theoretische Grundannahmen über den zu analysierenden Gegenstandsbereich zum Ausdruck. Demzufolge bestehen Unterschiede der *Kosten*, da nicht alle Schritte gleichwertig sind, sondern es kleine und große Schritte gibt, d.h. kleine und große Unterschiede zwischen den Zuständen: einige Zustände sind sich ähnlicher als andere. So kann z.B. in der Lebensverlaufsforschung dem Übergang von der Statusposition „vollzeitbeschäftigt“ zu „arbeitslos“ ein anderes Gewicht zukommen als der Wechsel von „vollzeitbeschäftigt“ zu „teilzeitbeschäftigt“.

Gewichtet werden können dabei grundsätzlich die *Kosten* für die Operation Substitution, sowie die *Kosten* für die Indel-Operationen. Durch die Definition dieser beiden Kostenarten definiert man gleichzeitig indirekt die Relation der Substitutionskosten zu den Indelkosten. Grundsätzlich können die *Kosten* einheitlich oder aber differenziert definiert werden.

Bei einer einheitlichen Kostendefinition wird jeweils *ein* Wert für die Operation „substitution“ und ein Wert für die Operationen „insertion“ und „deletion“ definiert. Als Standardeinstellung bzw. Grundeinstellung der Substitutionskosten wird die Definition einheitlicher Substitutionskosten von „1“ und einheitlicher Indelkosten von „0,5“ verwendet. Dies entspricht auch den Standard-Einstellungen von TDA (vgl. Rohwer / Pötter 2005; Brüderl / Scherer 2005: 6). Brüderl / Scherer (2005) weisen darauf hin, dass mit diesen Standardeinstellungen der einheitlichen Kostendefinition Sequenzen effektiv in Hinblick auf Muster und Strukturen analysiert werden können. Eine solche Analyse auf der Grundlage von Standardeinstellungen gilt ebenfalls für Verfahren wie die Clusteranalyse oder die Regressionsanalyse: Diese Verfahren werden in der Regel in der Standardeinstellung - wie sie z.B. in SPSS implementiert sind – genutzt, auch wenn eine differenziertere Verwendungsweise zu differenzierteren Ergebnissen führen könnte. Grundsätzlich ist jedoch nicht in jedem Fall eine solche Differenzierung notwendig (vgl. Brüderl / Scherer 2005: 13).

Für eine verfeinerte Analyse kann diese einheitliche Kostendefinition in Abhängigkeit der spezifischen Forschungsfrage jedoch weiter differenziert werden (Differenzierung der Kostendefinition). Diese bezieht sich

vor allem auf die differenzierte Definition der Substitutionskosten[51] in Form einer Substitutionskostenmatrix, wie sie im folgenden Kapitel ausgeführt wird.

6.4 Substitutions- und Indelkosten

Für die Optimal-Matching Analyse kommt der Operation der *Substitution* – und damit der Definition der Substitutions*kosten* – eine entscheidende Bedeutung zu, da es aus inhaltlich-interpretativer Perspektive einen bedeutenden Unterschied machen kann, ob z.B. der Zustand „vollzeiterwerbstätig“ durch „teilzeitbeschäftigt“ oder „arbeitslos“ ersetzt wird, um ein „alignment“ zwischen Sequenzen zu erreichen. Generell ist die differenzierte Gewichtung von Operationen durch *Kosten* ein zentraler Punkt der Sequenzanalyse. Mit dieser Gewichtung wird Einfluss darauf genommen, welche Sequenzen der Optimal-Matching Algorithmus als ähnlich identifiziert.

Bei der Definition der Substitutionskosten im Bereich der Sozialwissenschaften gilt die Devise von Abbott / Hrycak (1990: 155): „Big jumps cost more“ - je größer der Sprung desto höher die Substitutionskosten, je größer der Unterschied der substituierten Zustände, desto höher der Wert der Substitutionskosten. Unmittelbar daran schließt sich die Frage an, wie genau ein großer bzw. ein kleiner Sprung definiert ist? Welche Zustände ähneln sich wie stark im Vergleich zu anderen Zuständen?

Diese Frage nach der Gewichtung von Operationen durch *Kosten* kann also nur inhaltlich aufgrund theoretischer Überlegungen beantwortet werden und es wird deutlich, warum Abbott (1990a) davon spricht, dass der OM-Algorithmus als explorativ-heuristisches Verfahren mit einem *Minimum* an Vorannahmen auskommt. Dieses Minimum kommt genau in den theoretischen Annahmen zum Ausdruck, die Grundlage der Definition der Gewichtung von Operationen durch *Kosten* sind.

Auf diese Weise werden in der Lebensverlaufsforschung Übergänge zwischen Statuspositionen durch unterschiedliche Substitutionskosten berücksichtigt: „So kann das Austauschen einer Familienarbeit durch eine Vollzeitberufsarbeit 'teurer' – im Sinne von bedeutender – sein als die Ersetzung einer Teilzeitberufsarbeit durch eine berufsunabhängige Teilzeittätigkeit“ (Erzberger 2001: 150). Bei der Gewichtung von Operationen durch *Kosten* handelt es sich um eine *Schärfung* der OM-Analyse zur Identifizierung von Strukturen und Mustern, um ein Anpassen und Spezifizieren des Algorithmus an den Untersuchungsgegenstand: Wird ein Zustand durch einen anderen, ähnlichen Zustand ersetzt, werden für diesen Fall die Substitutionskosten niedriger angesetzt als bei einem weniger ähnlichen Zustand. Dies hat zur Folge, dass „die Distanz bei Se-

51 Eine differenzierte Definition von Indelkosten ist zwar grundsätzlich möglich (vgl. Sankoff / Kruskal 1999), jedoch programmtechnisch in TDA nicht implementiert. Die im Folgenden für die Definition der differenzierten Substitutionskosten ausgeführten Voraussetzungen einer theoretischen Fundierung gelten dabei grundsätzlich auch für die differenzierte Definition der Indelkosten.

quenzen die aus unterschiedlichen, aber sehr ähnlichen Zuständen bestehen“ (Brüderl / Scherer 2005: 5) kleiner wird.

Die Definition der Substitutionskosten erfordert dabei in Abhängigkeit des Zustandsraumes als Anzahl der unterschiedlichen Ereignisse ein komplexes Vorgehen. Da in einer Substitutionskostenmatrix für jede potenziell mögliche Kombination von Ereignissen (d.h. für jeden potentiell möglichen Übergang) die Substitutionskosten definiert sein müssen, ist ein entsprechend komplexes theoretisches Modell erforderlich, auf dessen Basis eine solche Definition vorgenommen werden kann.[52]

Die Definition von Substitutionskosten und das dazu erforderliche theoretische Modell stellt für Elzinga (2003, 2005) einen grundlegenden Ansatzpunkt der Kritik am Optimal-Matching Verfahren dar.[53] Das erforderliche theoretische Modell beschreibt er folgendermaßen: „Specifying a cost function embodies a very precise, numerical, (proto-) theoretical notion of the relative importance of the events that make up the sequence and the order in which they do or do not appear in a sequence. Therefore, the spatial representation of the sequences, as well as the metric of the space itself, is to be considered as a precise, geometrical model of a sociological theory“ (Elzinga 2003: 6). Elzingas Fazit lautet, dass in den Sozialwissenschaften ein derart detailliertes und reichhaltiges theoretisches Modell in der Regel nicht vorhanden ist: „Unfortunately, I do not know of sociological theories that permit such precise and rich geometrical models“ (Elzinga 2003: 6).[54]

Eine pragmatische Alternative zur differenzierten Definition von Substitutionskosten stellt neben der Verwendung der Standardeinstellungen (*default*) die datenbasierte Definition von Substitutionskosten dar, die TDA als eine Option der Durchführung der Optimal-Matching Analyse anbietet. Dabei werden die Substitutionskosten auf Grundlage der Übergangswahrscheinlichkeit als Häufigkeiten der Zustandswechsel im entsprechenden Datensatz berechnet. Grundannahme ist dabei, dass in den Daten häufig auftretende Übergänge spezifischer Zustände in der Empirie einfacher, d.h. mit weniger Kosten zu erreichen sind als weniger häufig auftretende Zustandswechsel.

Stellt beispielsweise der Übergang „A – B“ einen in den Sequenzen häufig enthaltenen Übergang dar, werden die Kosten der Substitution des Zustands „A“ durch den Zustand „B“[55] als geringer definiert als die

52 Die Schwierigkeiten der theoretischen Begründung der differenzierten Verwendung von Substitutionskosten wird auch bei Aisenbrey (2000: 27) deutlich. Sie verwendet zur beispielhaften Verdeutlichung der Substitutionskosten eine Matrix, die auf Schulnoten basiert. Die Kosten für die Substitution zweier Ereignissen (Noten) wird dabei definiert als die Differenz zwischen diesen Schulnoten. Ein Übergang von Note „1“ zu Note „2“ bedeutet eine Distanz von „1“; ein Übergang von Note „1“ zu Note „3“ bedeutet eine Distanz von „2“ usw. Die Indelkosten definiert Aisenbrey in Anlehnung an den maximal auftretenden Substitutionskosten mit dem Wert „3“ (in dem Beispieldatensatz ist die größte auftretenden Differenz die zwischen der Note „1“ und der Note „4“).

53 Zur kritischen Diskussion der Definition der Substitutionskosten, vgl. Wu (2000) und Levine (2000).

54 Diese Kritik an der Definition von Substitutionskosten stellt für Elzinga (2003) den Ausgangspunkt für die Entwicklung von Algorithmen zur Sequenzanalyse dar, die nicht auf den theoretischen Vorannahmen des Optimal-Matching Verfahrens basieren, sondern ausschließlich mit den konstitutiven Eigenschaften von Sequenzen arbeiten, d.h. den Elementen der Sequenz und deren Reihenfolge. Ähnlichkeit („Similarity“) wird demzufolge definiert als Anzahl gemeinsamer Vorgänger-Elemente: „Defining a precedence relation for a set of tokens is exactly what generates a sequence“ (Elzinga 2003: 11).

55 Aufgrund der Symmetrie der Substitutionskostenmatrix gilt dies auch für die Substitution des Zustandes „B“ durch den Zustand „A“.

Kosten für Übergänge, die selten in den Sequenzdaten enthalten sind. Konkret werden in TDA die errechneten Übergangswahrscheinlichkeiten vom dem als Standard für Substitutionskosten definierten Wert „2“ subtrahiert.[56]

Für die Optimal-Matching Analyse (und genauer: für die Optimum-Analyse als Berechnung der geringsten *Kosten* zur Herstellung eines „alignments“) bedeutet dies, dass in Abhängigkeit der Relation der Indel- und Substitutionskosten im konkreten Fall analysiert wird, ob die Verwendung der Operation Indel oder Substitution zu einer Übereinstimmung der Sequenzen mit den geringsten *Kosten* führt. Am Beispiel der Tabelle 2: 52 der datenbasiert errechneten Substitutionskosten ist damit ablesbar, dass

- sich die Substitutionskosten zwischen dem Wert „1“ und „2“ bewegen;
- die Substitution der Zustände „C“ und „B“ die geringsten Kosten (1.09043) verursacht und demzufolge als Übergang in den vorliegenden Sequenzen am häufigsten vorkommt;
- die Substitution der Zustände „H“ und „F“ die höchsten Kosten (1.97881) verursacht und demzufolge als Übergang in den vorliegenden Sequenzen am seltensten vorkommt;
- die höchsten Substitutionskosten knapp unter dem Wert von „2“ als *default*-Wert liegen für die am seltensten vorkommenden Übergänge in den Daten;
- die Matrix der Substitutionskosten generell symmetrisch aufgebaut ist, wobei die Diagonale von links oben nach rechts unten den Wert „0“ enthält.

Das Ergebnis dieser datenbasierten Berechnung der Substitutionskosten wird von TDA in Form einer Matrix ausgegeben und dokumentiert.

	A	B	C	D	E	F	G	H
A	0	1.89644	1.63055	1.88682	1.91262	1.92232	1.72264	1.23404
B	1.89644	0	1.09043	1.93872	1.64376	1.63187	1.94595	1.65869
C	1.63055	1.09043	0	1.96566	1.8628	1.94974	1.49433	1.51123
D	1.88682	1.93872	1.96566	0	1.16636	1.74983	1.45527	1.89392
E	1.91262	1.64376	1.8628	1.16636	0	1.22911	1.9258	1.92332
F	1.92232	1.63187	1.94974	1.74983	1.22911	0	1.9571	1.97881
G	1.72264	1.94595	1.49433	1.45527	1.9258	1.9571	0	1.82188
H	1.23404	1.65869	1.51123	1.89392	1.92332	1.97881	1.82188	0

Tabelle 2: Beispiel einer datenbasiert errechneten Substitutionskostenmatrix (TDA)

Analog zur Darstellung der Levenshtein-Distanzen in Form einer Matrix (Levenshtein-Distanzmatrix) werden auch die Substitutionskosten in einer Matrix (Substitutionskostenmatrix) dargestellt, in der systematisch alle potenziell möglichen Kombinationen von Ereignissen abgebildet werden. Wie die Levenshtein-Distanz-

56 Für eine detaillierte Beschreibung der datenbasierte Ermittlung der Substitutionskosten, vgl. Rohwer / Pötter (2005).

matrix ist auch die Substitutionskostenmatrix *symmetrisch* angelegt: Die Kosten für die Substitution des Elementes „A“ durch das Element „B“ entsprechen dabei den Kosten der Substitution des Elementes „B“ durch das Element „A“.

Das weiter oben erwähnte erste Distanzkonzept von Levenshtein ist aus der erweiterten Perspektive der Gewichtung von Operationen durch *Kosten* eine spezielle Variante, die darin besteht, dass die *Kosten* für alle Operationen durch den Wert „1“ definiert werden. Es wird also keine differenzierte Gewichtung vorgenommen und die Anzahl der minimalen Operationen bestimmt die Levenshtein-Distanz. Diese spezielle Variante liegt auch der Abbildung 9: 46 zugrunde. In die Ecken der einzelnen Zellen sind jeweils die *Kosten* unterschiedlichen Operationen (Einfügen, Löschen, Austauschen) eingetragen. In dem von Erzberger (2001) verwendeten Beispiel sind die Kosten für alle drei Operationen einheitlich mit „1“ definiert.

Wird jedoch eine differenzierte Gewichtung von Operationen vorgenommen, werden diese differenzierten Kosten für die Operationen in die Zellen der Matrix zur Ermittlung der Levenshtein-Distanz eingetragen. Für den Fall der Substitution eines Elementes durch ein anderes wird dann anhand der Substitutionsmatrix der entsprechende Wert ermittelt und in die Tabelle eingetragen. Für die oben dargestellte Matrix zur Ermittlung der Levenshtein-Distanz (Abb. 9:46) bedeutet dies, dass der Wert „1“ in der linken oberen Ecke der Zellen der Matrix durch die jeweiligen Werte für die spezifische Substitution ersetzt würde (z.B. den Wert für die Substitution des Zustandes „A“ durch den Zustand „B“; des Zustandes „C“ durch den Zustand „A“ usw.).

6.5 Relation Substitutionskosten - Indelkosten

Im vorangehenden Abschnitt wurde die differenzierte Gewichtung der Operation *Austauschen* („substitution“) beschrieben. Aufgrund der bereits erwähnten Komplementarität der Indel-Operationen werden für die Operationen „insertion“ und „deletion“ grundsätzlich die gleichen Kosten angesetzt. In diesem Kapitel wird nun die Gewichtung der Indel-Operationen näher erläutert, d.h. vor allem die Relation der Indelkosten zu den Substitutionskosten. Analog der Spezifizierung der Definition der Substitutionskosten stellt auch die Definition der Indelkosten eine Schärfung und Anpassung des OM-Algorithmus an den Untersuchungsgegenstand dar. Generell werden folgende Möglichkeiten der Definition der Indelkosten verwendet (vgl. Aisenbrey 2000: 29):

- Die Indelkosten werden an den *maximalen Substitutionskosten* ausgerichtet (z.B. Abbott / Hrycak 1990; Stovel / Savage / Bearman 1996, Aisenbrey 2000).
- Die Indelkosten werden an den *minimalen Substitutionskosten* ausgerichtet (z.B. Erzberger / Prein 1997).

Bei der Definition der Relation zwischen Substitutions- und Indelkosten ist grundsätzlich die Vorgehensweise des Optimal-Matching Algorithmus sowie die Kosten-Minimierung zu berücksichtigen. Ausgehend vom Standard (*default*) einheitlicher Substitutionskosten von „2“ und der Indelkosten von „1“ als Grundeinstellung von TDA ergeben sich folgende Überlegungen (vgl. Brüderl / Scherer 2005):

- Werden die Kosten für die Operation „substitution“ höher „2“ definiert (wobei „2“ der Summe der Kosten für eine „insertion“ und eine „deletion“-Operation entspricht), wird bei der Errechnung der minimalen Kosten die Operation „substitution“ nicht mehr berücksichtigt, da diese im Vergleich zu den Indelkosten zu „teuer“ ist, d.h. zu höheren Kosten führt als die Verwendung der Operationen „insertion“ und „deletion“. Die Verwendung dieser Operationen führt in diesem Fall immer zu geringeren Kosten als die Verwendung der Operation „substitution“. Diese Definition der Relation der Substitutions- und Indelkosten entspricht dann der zweiten Variante des Distanzmaßes von Levenshtein.
- Werden die Kosten für die Operation „substitution“ wesentlich geringer als „2“ definiert (wobei „2“ der Summe der Kosten einer „insertion“ und einer „deletion“-Operation entspricht), werden analog zum obigen Fall bei der Errechnung der minimalen Kosten die Operationen „insertion“ und „deletion“ nicht mehr verwendet, da diese im Vergleich zu den Substitutionskosten zu „teuer“ sind. Diese Definition der Relation von Substitutions- und Indelkosten entspricht dann der Hamming-Distanz als einfachem Substitutionsalgorithmus.

So weist auch Erzberger (2001: 150) darauf hin, dass die *Relation von Indel-* und *Substitutionskosten* darüber entscheidet, ob Sequenzen durch *Einfügen* oder *Löschen* gegeneinander verschoben werden. „Ein niedrigerer 'Kostenwert' dieser [Indel, S.I.] Operationen gibt dem Vorgang des 'Löschens' und 'Einfügens' Vorrang vor dem des 'Ersetzens'. Damit ist bei gegeneinander verschobenen Verläufen die Aktion des Verschiebens des Zielverlaufes 'billiger' als das Ersetzen. Ein hoher [Indel, S.I.] Wert dagegen nimmt auf eine gegenseitige Verschiebung keine Rücksicht“ (Erzberger 2001: 150).

Generell ist die Definition der differenzierten Kosten abhängig von der Forschungsfrage und nicht allgemeingültig zu beantworten. Konkret definiert Erzberger (2001: 152) die Substitutionskosten mit „1“ und die Indelkosten mit „0,5“,[57] um den Indel-Operationen Vorrang vor den Substitutionen zu geben und somit das „alignment“ durch gegenseitiges Verschieben billiger zu machen und damit Sequenzen als ähnlich zu definieren, die durch *Einfügen* und *Löschen* ineinander überführt werden können. In gleicher Zielrichtung schlagen Brüderl / Scherer (2005: 6) vor, die Kosten für die Summe der Indel-Operationen gleich den Substitutionskosten (bzw. etwas darunter) anzusiedeln: „Bei Indelkosten von 1 sollten sich also die Substitutionskosten zwischen 1 und 2 bewegen“. Die Orientierung einheitlicher Indelkosten an den minimalen bzw. maxima-

57 Diese Definition entspricht einem Verhältnis von 2 zu 1.

len Substitutionskosten stellen dabei die Pole als *Definitionsspielraum* der differenzierten Kostendefinition dar.

Auf ein alternatives forschungspragmatisches Verfahren zur Definition der Relation von Substitutions- und Indelkosten wurde bereits im vorangegangenen Kapitel unter Hinweis auf eine datenbasierte Berechnung der Substitutionskosten aus den Übergangswahrscheinlichkeiten hingewiesen. Ausgangspunkt dafür ist die Verwendung der Indelkosten mit „1“ und der datenbasierten Errechnung der Substitutionskosten. Dieses Vorgehen wird von Rohwer / Trappe (1997) verwendet. Die Substitutionskosten liegen dabei in Abhängigkeit der Übergangswahrscheinlichkeiten durchgängig unter dem Wert der Summe der Operationen für „insertion“ und „deletion“ (vgl. Rohwer / Pötter 2005: VIII-60).

6.6 Sequenzen unterschiedlicher Länge

Grundsätzlich ist der Optimal-Matching Algorithmus in der Lage, unterschiedlich lange Sequenzen zu vergleichen, d.h. Sequenzen mit einer unterschiedlichen Anzahl von Elementen. Auf den Einfluss der Sequenzlänge auf die Anzahl der Operationen und damit auf die Kosten zum Erreichen einer Übereinstimmung („alignment“), sowie auf die Möglichkeit der Berechnung des *relativen Distanzmaßes* als Verhältnis von Transformationskosten zur Sequenzlänge wurde bereits in Kapitel 6.5 hingewiesen.

In Abhängigkeit des Forschungsgegenstandes und vor allem dann, wenn unterschiedlich lange Sequenzen gleichzeitig unterschiedliche Beobachtungszeiträume darstellen, wie z.B. in der Lebensverlaufsforschung, kann es sinnvoll sein, unterschiedlich lange Sequenzen vor der Optimal-Matching Analyse zu *standardisieren*, um eine Vergleichbarkeit überhaupt erst herzustellen. Im Rahmen der soziologischen Ereignisdatenanalyse spricht man in diesem Zusammenhang von *Zensierung* („censoring“) der Sequenzen und unterscheidet verschiedene Arten (vgl. Blossfeld / Rohwer 2002). In Abbildung 10 wird auf der x-Achse die (historische) Zeit abgetragen, der Beobachtungszeitraum ist in der Regel ein Ausschnitt mit definierter zeitlicher Länge.

Grundsätzlich wird eine *Links*- von einer *Rechts*zensierung unterschieden, wobei sich diese Begriffe auf das *Beobachtungsfenster* („observation window“) beziehen. So wird beispielsweise eine Linkszensierung in der Regel dann durchgeführt, wenn die Aufzeichnung der Daten über die Aufenthaltsdauer im Ausgangszustand unvollständig ist (vgl. Blossfeld / Rohwer 2002: 39). An dieser Stelle können nicht die Implikationen der unterschiedlichen Zensierungsarten dargestellt werden. Vielmehr soll anhand eines Vergleiches das Vorgehen der Navigationsanalyse näher erläutert werden.

Ein grundlegender Unterschied besteht zwischen den Daten der Ereignisdatenanalyse (vgl. Kap. 8) und den Daten der Navigationsanalyse als Sequenzanalyse: Bei der Ereignisdatenanalyse stellt der Lebenslauf die zeitliche Achse dar, wobei die lebenslaufspezifischen Daten in der Regel retrospektiv erhoben werden. Der

Beobachtungszeitraum der in Kapitel 9 durchgeführten Navigationsanalyse beginnt mit der Anmeldung an der Lernplattform und endet mit deren Verlassen. Die Daten über den Navigationsverlauf als Abfolge aufgerufener Seiten und der Verweildauer auf diesen Seiten werden *während* des Prozesses über Logfiles automatisiert aufgezeichnet (periaktional, vgl. Kap. 2).

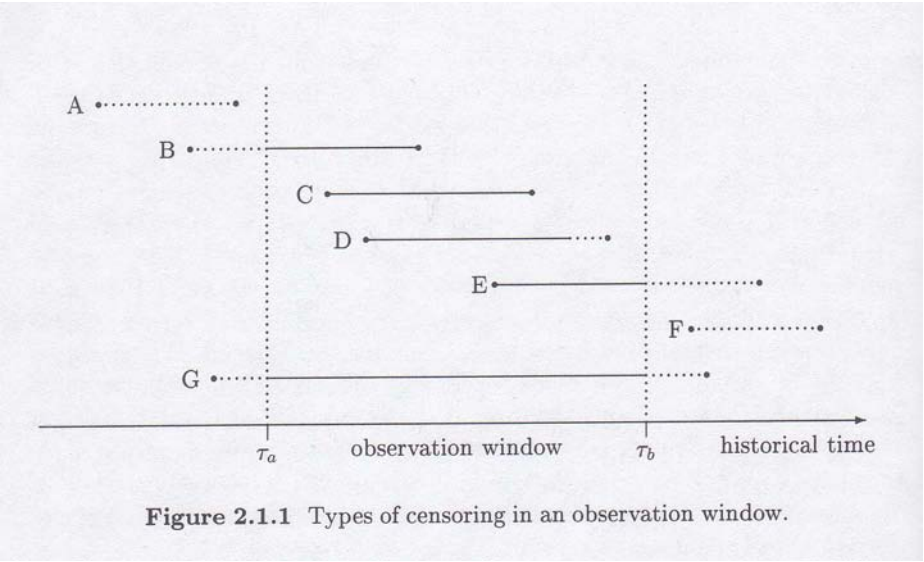

Figure 2.1.1 Types of censoring in an observation window.

Abbildung 10: Arten der Zensierung von Sequenzen (Blossfeld / Rohwer 2002: 40)

Neben den oben beschriebenen Arten der Zensierung von Sequenzen bestehen nach Aisenbrey (2000: 28) im Rahmen der Sequenzdatenanalyse folgende grundsätzlichen Möglichkeiten:

- Es werden unterschiedliche Sequenzlängen verwendet (vgl. Abbott / Hrycak 1990; Stovel et al. 1996);
- die Länge der Sequenzen (d.h. die Anzahl der Elemente) beim paarweisen Vergleich wird standardisiert, in dem die Sequenzlänge auf die Anzahl der Elemente der kürzeren Sequenz reduziert wird. Bei der längeren der zu vergleichenden Sequenzen werden alle Elemente *zensiert*, die über die Länge der kürzeren Sequenz hinausgehen (vgl. Erzberger / Prein 1997);
- die Sequenzlänge wird bereits bei der Datenerhebung standardisiert, z.B. durch die Definition des Zeitfensters, so dass alle zu analysierenden Sequenzen die gleiche Länge aufweisen (vgl. Halpin / Chan 1998).

Im Fall der beispielhaften Analyse von *Mikronavigationsprozessen* in Kapitel 7, *Sequenzanalyse am Beispiel* wird das Beobachtungsfenster zeitlich variabel definiert über den Prozess des Anmeldens bzw. des Verlassens spezifischer Lerneinheiten: Linkszensiert wird der Navigationsverlauf, der bis zu dieser spezifischen Lerneinheit geführt hat (also alle Lerneinheiten, die vor der analysierten Lerneinheit besucht wurden); rechtszensiert wird der Navigationsverlauf, der auf den Besuch der analysierten Lerneinheit folgte. Denn

dies bedeutet gerade die Fokussierung auf Prozesse der Mikronavigation: Aus analytischen Gründen werden Prozesse der *Mikro*navigation aus den der *Makro*navigation herausgelöst. Die in diesem Beobachtungsfenster erhobenen Mikronavigationsprozesse weisen unterschiedliche Längen sowie unterschiedliche Abfolgen von Wissenseinheiten auf und sind grundsätzlich als „multistate-multi-episode process“ (Blossfeld / Rohwer 2002: 39) zu interpretieren.

Im Rahmen des explorativ-heuristischen Charakters der Navigationsanalyse ist darüber hinaus eine Standardisierung auf eine einheitliche zu analysierende Sequenzlänge nicht wünschenswert, da Unterschiede im Navigationsverhalten gerade in der Auswahl und der Anzahl von Wissenseinheiten zum Ausdruck kommen. Die Länge einer spezifischen Navigationssequenz ist somit Ausdruck einer spezifischen Bearbeitungsweise.

Im Gegensatz zur Definition einer Sequenz als aus mindestens zwei Elementen bestehend (vgl. Kap. 5.1: 34) werden bei der Navigationsanalyse auch „Einer“, d.h. einmalige Aufrufe einer Wissenseinheit innerhalb der zu analysierenden Lerneinheit berücksichtigt. Bei diesen einmaligen Aufrufen handelt es sich in der Regel um die Wissenseinheit *Orientierung*, da diese automatisch als erste Wissenseinheit der übergeordneten Lerneinheit angezeigt wird, als *didaktische Station* (vgl. Kap. 3.1; *Grundlagen der Web-Didaktik*). Diese „Einer“ werden als Hinweis auf eine der Mikronavigation übergeordnete Makronavigation interpretiert.

6.7 Potential der Optimal-Matching Analyse

In diesem Abschnitt wird das analytische Potential der Sequenzanalyse mittels Optimal-Matching diskutiert. Den Ausgangspunkt bildet die Argumentation Abbotts (Abbott/ Forrest 1986, Abbott 1990a, Abbott 1990b; Abbott 1990c; Abbott / Hrycak 1990, Abbott1992, Abbott 1995a, Abbott / Barman 1997, Abbott / Tsay 2000), insbesondere seine Kritik an den bis dato verwendeten Verfahren der Analyse von Verlaufs- und Sequenzdaten.

Allgemein grenzt Abbott die Sequenzdatenanalyse auf Grundlage des Optimal-Matching von am Querschnittdesign orientierten Analyseverfahren ab: Gegenstand der Sequenzdatenanalyse sind Sequenzen als zeitlich zusammenhängende *Abfolgen* bzw. *Ketten* und ausdrücklich nicht die Analyse *einzelner, aggregierter* Ereignisse oder *isolierter* Übergänge. In dieser Perspektive handelt es sich bei der Aggregation um eine unzulässige Reduktion der Komplexität von Sequenzen als zeitlicher Abfolgen auf dyadische Abfolgen, ohne die Sequenz als solche zum Gegenstand der Analyse zu machen. In methodischer Hinsicht besteht der Hauptkritikpunkt an letzterem Vorgehen in der fehlenden Berücksichtigung gerade der zeitlichen Information, die eine Sequenz ausmacht.

Bei der Fokussierung auf die Analyse und Erklärung von Sequenzen als zeitlichen Prozessen unterscheidet Abbott (1990a) „pattern questions“ (Gibt es gemeinsame Muster in Sequenzen?) von „generation questions“

(Wenn es Muster gibt, wie werden diese generiert?), also Fragen des Vorhandenseins von Mustern von Fragen der Entstehung und Entwicklung dieser Muster.

Abbotts Kritik bezieht sich in erster Linie auf das Ausblenden der „pattern questions“ und die Fokussierung der „generation questions“. Abbott kritisiert, dass damit Sequenzen als solche überhaupt nicht zum Gegenstand der Analyse werden: „In practice, the pattern question has proved so problematic that the most common strategy with sequences has been to answer the generation question with an estimated model and then to see wether the model generates sequences whose aggregated properties (not including sequence patterns) resemble those of the original data“ (Abbott / Hrycak 1990: 148). Abbott zufolge sind das Ausblenden sowie die beschriebene Fokussierung vor allem auf das Fehlen geeigneter Methode der Analyse von Sequenzdaten zurückzuführen.

Pointiert formuliert kritisiert also Abbott, dass die Frage nach dem Vorhandensein gemeinsamer Muster in Sequenzen systematisch ausgeblendet wird und darüber hinaus die Frage nach der Entstehung und Entwicklung von Mustern auf der Grundlage *aggregierter* Daten beantwortet wird. Diese können jedoch seiner Meinung nach aus methodologischen Gründen keinerlei Beitrag zur Beantwortung dieser Frage leisten, denn Sequenzdaten werden dabei aggregiert und als aggregierte Daten analysiert. Zur Entstehung und Entwicklung von Mustern wird ein Modell entworfen, dessen Erklärungspotential wiederum anhand aggregierter Daten überprüft wird. Kurz: Die Analyse von Sequenzen erfolgt jeweils über den Zwischenschritt der Aggregation.

Anstelle der Entwicklung eines Modells zu Erklärung von Mustern und in Ablehnung der Analyse von Sequenzen anhand aggregierter Daten schlägt Abbott die Methode der Sequenzdatenanalyse auf Grundlage der Optimal-Matching Analyse vor: Die *Analyseeinheiten* bilden dabei gerade Sequenzen als Abfolge *zeitlich zusammenhängender* Zustände und nicht aggregierte Querschnittsdaten (z.B. als Kennzahlen in Form der Verweildauer, Anzahl der Zustände oder der Zustandswechsel). Insofern kann die Sequenz als Kontext der einzelnen, sie definierenden Ereignisse angesehen werden.

Generell bezieht sich Abbott in seinen Ausführungen auf die *quantitative Analyse* von Sequenzdaten auf der Grundlage der Verwendung des Optimal-Matching Algorithmus - und nicht auf die *qualitative Analyse,* z.B. als Einzelfallanalyse. Übergeordnetes Ziel ist der Vergleich einer großen Anzahl komplexer und oftmals sehr langer Sequenzen, das *Identifizieren von Mustern, Strukturen und Regelmäßigkeiten* in diesen Sequenzen. Dieses Identifizieren ist ohne die Verwendung von Computertechnologie und spezifischer Algorithmen nicht möglich. Offensichtlich wird die Leistungsfähigkeit der Sequenzanalyse am Beispiel der Analyse der extrem komplexen und komplizierten Struktur der DNA.

Grundsätzlich schafft die Optimal-Matching Analyse die Ausgangsbasis für sowohl eine *explorativ-heuristische* als auch eine *konfirmatorische* Forschungsstrategie, „[...] sequence comparison is involved in drawing up questions as well as in answering them“ (vgl. Kruskal 1999: 31). Bei diesen Strategien wird direkt auf die Levensthein-Distanzmatrix als Ergebnis der Optimal-Matching Analyse zugegriffen, jedoch in unterschiedlicher Weise (vgl. Kap. 5.2, 36f.)

In einer *konfirmatorischen* Forschungsstrategie werden theoretische Modelle oder prototypische Sequenzen empirisch vorhandenen Sequenzen gegenüber gestellt (vgl. Erzberger 2001, Chan 1995). Empirische vorhandene Sequenzen werden dabei mit im Vorhinein – auf theoretischer Grundlage - gebildeten Sequenzen (z.B. „Idealtypen“) als Referenzsequenzen verglichen. Über einen solchen Vergleich können einerseits empirische Verläufe unterschiedlichen theoretischen Typen zugeordnet werden sowie andererseits das empirische Vorhandensein theoretischer Sequenzen überprüft werden[58] Innerhalb multivariater Analysemethoden verorten Backhaus et al. (vgl. 2000: XXI) dieses Vorgehen als primär *strukturen-prüfendes* Verfahren. Im Rahmen der Navigationsanalyse können somit z.B. empirisch vorhandene Sequenzen mit typischen, aus der Theorie entwickelten Sequenzen (z.B. Navigationsstrategien, Navigationstypen, Lernstrategien) verglichen werden.

Bei der *explorativ-heuristischen* Forschungsstrategie werden Sequenzen als empirisch erhobenen Daten in einem ersten Schritt miteinander unter dem Gesichtspunkt der Distanz verglichen und in einem zweiten Schritt mit Hilfe von Methoden der Clusteranalyse zu Gruppen ähnlicher Sequenzen zusammengefasst. Innerhalb multivariater Analysemethoden verorten Backhaus et al. (vgl. 2000: XXI) dieses Vorgehen als primär s*trukturen-entdeckendes* Verfahren.

Das Vorgehen der Sequenzdatenanalyse im Rahmen der Navigationsanalyse entspricht einer explorativ-heuristischen Strategie. Mit Erzberger (2001) und Baur (2005) kann diese explorativ-heuristische Sequenzdatenanalyse als *fallorientierte Analysestrategie* gekennzeichnet werden, bei der Sequenzen als Gesamtverläufe bzw. Verlaufsgeschichten in ihrer Vielfalt und Komplexität zum Gegenstand der Forschung werden.[59] Dabei gehen die einzelnen Sequenzen der untersuchten Fälle in ihrer Gesamtheit in die Analyse ein und die „Zusammenschau aller Verläufe läßt dann Ordnung entstehen“ (Erzberger 2001: 136). In der Gesamtschau einer hinreichend großen Anzahl von Sequenzen werden spezifische Muster oder Regelmäßigkeiten überhaupt erst erkennbar (vgl. Abbott 1990).

Die Sequenzdatenanalyse als heuristisches und exploratives Verfahren (vgl. Erzberger / Prein 1997) ermöglicht es, „typische Muster, die sich aus der Empirie ergeben, theoretisch aber nicht 'vorgedacht' wurden“ (Aisenbrey 2000: 15) zu identifizieren. Dies ist ein wesentlicher Unterschied zu konfirmatorischen Strategien und zu dem Verfahren der Ereignisdatenanalyse. Bei der *konfirmatorischen Analyse* empirischer Sequenzen anhand im Voraus definierter (theoretischer) Muster wird das Vorhandensein dieser theoretischen Muster überprüft bzw. der Grad der Abweichung (Distanz) der empirischen von den theoretischen Sequenzen. Muster in empirischen Sequenzen, die theoretisch nicht 'vorgedacht' wurden, können durch eine konfirmatorische Analyse jedoch nicht identifiziert werden und kommen lediglich als *Abweichungen* des überprüften theoretischen Modells in den Blick (ohne allerdings Gegenstand genauerer Beschreibung oder Analyse zu werden).

58 Bei der konfirmatorischen Forschungsstrategie wird insbesondere auf Methoden der Clusterung verzichtet. Vermieden werden damit potentielle Fehlerquellen der Clusteranalyse, die z.B. durch die Auswahl eines spezifischen Clusteralgorithmus oder der Definition der Clusteranzahl entstehen können (vgl. Kap. 9: 96).

59 In der Terminologie von Baur (2005: 113ff.) entspricht die fallbezogene Betrachtungsweise einer *zeilenweisen* und die variablenbezogene Betrachtungsweise einer *spaltenweisen* Analyse der *Ereignismatrix*.

Die Bedeutung der Sequenzdatenanalyse mittels Optimal-Matching bezeichnet Erzberger (2001: 135) als eine „Neuorientierung der quantitativen Längsschnittforschung: weg von einer überprüfenden, auf erklärenden Variablen fußenden Analyse und hin zu Verfahren, die fallorientiert mit einem Minimum ex ante getätigten Festlegungen arbeiten.“ Forschungsmethodisch steht dabei eine Perspektive im Vordergrund, „die sich nicht in der Feststellung eines zu einem Zeitpunkt 'x' gemessenen *Zustandes* erschöpft, sondern an dem *Weg* interessiert ist, der zu diesem Zustand führt“ (Erzberger 2001: 135). Generell ist das explorativ-heuristische Vorgehen eher dem *context of discovery* zuzuordnen, das konfirmatorische Vorgehen eher dem *context of justification* (vgl. Erzberger 2001).

In dieser Perspektive kann die Sequenzanalyse als *deskriptives* Verfahren (vgl. Elzinga 2003) bezeichnet werden, das Sequenzen ohne den Rückgriff auf *aggregierte Daten* analysiert und dabei ohne ein im Vorhinein (ex ante) definiertes theoretisches Modell der Entstehung oder des Vorhandenseins von Mustern arbeitet. Grundsätzlich schließt die Methodologie der explorativ-heuristischen Sequenzanalyse, wie sie von Abbott vorgeschlagen wird, jedoch nicht die Entwicklung und Überprüfung theoretischer Modelle aus. Der Entwurf eines theoretischen Modells stellt einen über die Optimal-Matching Analyse hinausgehenden Schritt dar, für den die explorativ-heuristische Analyse empirischer Sequenzen eine empirische Grundlage darstellen kann. So können die Ergebnisse der Sequenzdatenanalyse - z.B. in Form der Clusterzugehörigkeit einzelner Sequenzen - in weiteren Analyseschritten innerhalb hypothesenprüfender Verfahren als abhängige bzw. unabhängige Variable verwendet werden. Diese weiteren Analyseschritte bauen auf dem Ergebnis der Optimal-Matching Analyse auf, sind jedoch nicht mehr explizit Gegenstand der Optimal-Matching Analyse.

7 Sequenzanalyse am Beispiel

Zur Vertiefung und zur Verdeutlichung der Funktionsweise der Sequenzdatenanalyse mittels Optimal-Matching wird in diesem Kapitel beispielhaft ein überschaubarer, fiktiver Datensatz analysiert. Folgende 11 Sequenzen bilden dabei den Ausgangspunkt: Sie bestehen jeweils aus den gleichen fünf Elementen, wobei jeweils 3 Sequenzen identisch sind (vgl. Abbildung 11: 61).

Analysiert werden diese Sequenzen mittels Optimal-Matching, wobei unterschiedliche Definitionen der Substitutionskosten (vgl. Kap. 6.4: 50) miteinander verglichen werden: *default*-Substitutionskosten, *datenbasierte* Substitutionskosten und Substitutionskosten als *absolute Differenz*.

Daran anschließend werden die Sequenzen auf Grundlage der berechneten Levenshtein-Distanzmatrix durch das Ward-Clusterverfahren in möglichst homogene Gruppen eingeteilt und diese Ergebnisse zusammenfassend interpretiert.

```
1)      1 2 3 4 5
2)      1 2 3 4 5
3)      1 2 3 4 5

4)      2 1 3 4 5
5)      2 1 3 4 5
6)      2 1 3 4 5

7)      1 4 3 2 5
8)      1 4 3 2 5
9)      1 4 3 2 5

10)     5 4 3 2 1

11)     1 3 5 4 2
```

Abbildung 11: Datensatz mit 11 beispielhaften Sequenzen

Analysiert man diese Sequenzdaten mit dem Programm TDA mittels Optimal-Matching (seqm) erhält man eine Ausgabedatei („output-file“), in der das konkrete Vorgehen detailliert dokumentiert ist. Diese Dokumentation enthält Informationen darüber, ob die Optimal-Matching Analyse erfolgreich ausgeführt wurde. Im Fall einer nicht erfolgreichen Durchführung kann die exakte Stelle identifiziert werden, an der ein Fehler aufgetreten ist oder der Prozess abgebrochen wurde.

7.1 *Default*-Substitutionskosten

Im Folgenden wird zur Veranschaulichung und Erläuterung die vollständige Ausgabedatei abgebildet. Ausgangspunkt ist eine Optimal-Matching Analyse auf der Grundlage der *default*-Definition der *Substitutionskosten* (vgl. 6.4: 50; vgl. auch Zeile 85, 86).

```
1. TDA. Analysis of Transition Data (6.4k). Fri May 19 16:28:07 2006
2. Current memory: 330832 bytes.
3.
4. Reading command file: D:\_Diss\Diss_Navigation\logdaten\grep\testwien\testwien.cf
5. ================================================================================
6. rspss1(...)=D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq01.sav
7. Reading SPSS sav file: D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq01.sav
8.
9. Identification: $FL2@(#) SPSS DATA FILE MS Windows Release 12.0 spssio32.dll
10.Number of OBS elements per observation: 6
11.Compression switch: 1
12.Index of case-weight variable: 0
13.Number of cases: 11
14.Compression bias: 100
15.
16.Creation date: 19 May 06
17.Creation time: 16:27:34
18.File label:
19.
20.Number of variables: 6
21.Number of string variables: 0
22.
23.Reading data to check variables.
24.Read 11 records.
25.Number of blank-type missing values: 0
26.Number of system-type missing values: 0
27.
28.Idx Variable  T   S  PFmt  Definition
29.-----------------------------------
30.  1 NUM       3   1   2.0  spss(0)
31.  2 V1        3   1   2.0  spss(0)
32.  3 V2        3   1   2.0  spss(0)
33.  4 V3        3   1   2.0  spss(0)
34.  5 V4        3   1   2.0  spss(0)
35.  6 V5        3   1   2.0  spss(0)
36.
37.Reading data again to create internal data matrix.
38.Maximum number of cases: 11
39.Allocated 66 bytes for data matrix.
40.
41.Read 11 records.
42.Created a data matrix with 6 variables and 11 cases.
43.-------------------------------------------------------------------------------
44.nvar(...)
45.Creating new variables. Current memory: 331037 bytes.
46.
47.Idx Variable  T   S  PFmt  Definition
```

```
48.-----------------------------------
49.  1 ID        3   4   0.0  NUM
50.  2 Y0        3   4   0.0  V1
51.  3 Y1        3   4   0.0  V2
52.  4 Y2        3   4   0.0  V3
53.  5 Y3        3   4   0.0  V4
54.  6 Y4        3   4   0.0  V5
55.
56.New variables will be added to existing data matrix.
57.Trivial matching.
58.
59.Added 6 variable(s) to existing data matrix.
60.Number of cases with no match: 0
61.
62.End of creating new variables. Current memory: 331410 bytes.
63.------------------------------------------------------------------------------
64.seqdef=Y0,,Y4
65.Creating a new sequence data structure. Current memory: 331410 bytes.
66.Sequence structure number: 1
67.Sequence type: 1
68.Currently defined sequences:
69.
70.Sequence          State       Time axis       Number
71.Structure Type  Variables  Minimum  Maximum  of States  States
72.-------------------------------------------------------------
73.     1      1         5          0        4         5    1 2 3 4 5
74.
75.Range of common time axis: 0 to 4.
76.------------------------------------------------------------------------------
77.seqm(...)=D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq_default.df
78.Sequence proximity measures. Current memory: 331442 bytes.
79.Optimal matching.
80.Using sequence data structure 1.
81.Number of states: 5. Max sequence length: 5
82.Option (sm=2): skip identical states.
83.Test output will be written to: wienseq01_default.tst
84.
85.Default indel cost: 1.
86.Default substitution cost: 2.
87.
88.
89.Starting alignment procedure.
90.Number of sequences (cases): 11
91.Sequences with zero length or internal gaps: 0
92.Sequences used for alignment: 11
93.
94.Number of alignments: 55
95.55 record(s) written to output file:
  D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq_default.df
96.Maximum distance between sequences 10 and 1: 8
97.------------------------------------------------------------------------------
98.Current memory: 330832 bytes. Max memory used: 331766 bytes.
99.End of program. Fri May 19 16:28:07 2006
100.
```

In der Ausgabedatei werden folgenden Informationen bzw. Schritte dokumentiert:

- In Zeile 4 der Name der gelesenen Befehlsdatei (command file, *.cf);
- in Zeile 11 die Anzahl der berücksichtigten Fälle;
- in Zeile 20 die Anzahl der unterschiedlichen Elemente der Sequenzen;
- in Zeile 25 die Anzahl von Lücken („gaps“ bzw. „missing values“);
- in Zeile 35 die Erstellung einer internen Datenmatrix ;
- in Zeile 44 die Erstellung neuer Variablen für diese Datenmatrix;
- in Zeile 64 die Erstellung einer neuen Sequenzdatenstruktur;
- in Zeile 73 grundlegende Eigenschaften der Sequenzen, z.B. die Anzahl unterschiedlicher Elemente und die Namen der unterschiedlichen Elementen;
- in Zeile 79 die Durchführung der Optimal-Matching Analyse (seqm);
- in Zeile 81 die maximale Sequenzlänge;
- in Zeile 82 die Art des Umgangs mit identischen Elementen;
- in Zeile 83 die Ausgabe einer Dokumentationsdatei (*.tst);
- in Zeile 85 die verwendeten Indelkosten;
- in Zeile 86 die verwendeten Substitutionskosten;
- in Zeile 89 der Beginn des „alignment“-Verfahrens;
- in Zeile 90 die Anzahl der gelesenen Sequenzen;
- in Zeile 91 die Anzahl der Sequenzen ohne Elemente bzw. mit internen Lücken;
- in Zeile 92 die Anzahl der verwendeten Sequenzen;
- in Zeile 94 die Anzahl der durchgeführten „alignments“;
- in Zeile 95 die Ausgabe einer Datei, die das Ergebnis des „alignments“ enthält (datafile, *.df);
- in Zeile 96 die Sequenzen mit der maximalen Distanz.

Im Folgenden wird die in Zeile 83 dokumentierte Ausgabedatei (*.tst) dargestellt. Sie enthält neben Informationen über die Anzahl der Elemente, der maximalen Sequenzlänge, der Indelkosten und der Substitutionskostenmatrix als zusätzliche Option Informationen über die Durchführung des „alignment“-Prozesses: Für jeden paarweisen Sequenzvergleich wird die entsprechende Matrix zur Ermittlung der Levenshtein-Distanz dokumentiert („D Matrix“, Distanzmatrix).[60] Da der hier zu Grunde gelegte Satz von Beispielsequenzen identische Sequenzen enthält, wird eine um diese Doppelungen gekürzte Darstellung gewählt, die gekürzten Stellen sind durch drei Punkte in eckigen Klammern „[...]“ gekennzeichnet.

Die Ausgabedatei enthält zwar die entsprechende Matrix zur Ermittlung der jeweiligen Levenshtein-Distanz, jedoch keine Information über den konkreten Weg, der zur Ermittlung der geringsten Distanz führt (vgl. 9:

60 In der TDA-Syntax wird diese zusätzliche Option innerhalb des Befehls *seqm* durch den Parameter *tst=3* umgesetzt.

46). Dieser Weg wurde nachträglich in das Ausgabedokument in Form einer Fettformatierung der Zellen der Matrizen eingefügt (entsprechend der Markierung durch Pfeile in Abbildung 6.2: 46). Darüber hinaus wurde aus Gründen der Übersichtlichkeit eine Nummerierung der Matrizen (a – j) eingefügt. Abschließend wird das konkrete Vorgehen zur Ermittlung der Levenshtein-Distanz näher beschrieben.

```
Optimal matching test output file.
Number of states: 5
Max sequence lenght: 5

Indel cost
1 1 1 1 1

Substitution cost
0 2 2 2 2
2 0 2 2 2
2 2 0 2 2
2 2 2 0 2
2 2 2 2 0

[...]
```

a)

D Matrix		0	1	2	3	4	5
		B	1	2	3	4	5
0	A	**0.00**	1.00	2.00	3.00	4.00	5.00
1	2	**1.00**	2.00	1.00	2.00	3.00	4.00
2	1	2.00	**1.00**	**2.00**	3.00	4.00	5.00
3	3	3.00	2.00	3.00	**2.00**	3.00	4.00
4	4	4.00	3.00	4.00	3.00	**2.00**	3.00
5	5	5.00	4.00	5.00	4.00	3.00	**2.00**

```
[...]
```

b)

D Matrix		0	1	2	3	4	5
		B	1	2	3	4	5
0	A	**0.00**	1.00	2.00	3.00	4.00	5.00
1	1	1.00	**0.00**	1.00	2.00	3.00	4.00
2	4	2.00	1.00	**2.00**	3.00	2.00	3.00
3	3	3.00	2.00	3.00	**2.00**	3.00	4.00
4	2	4.00	3.00	2.00	3.00	**4.00**	5.00
5	5	5.00	4.00	3.00	4.00	5.00	**4.00**

```
[...]
```

c)

D Matrix		0	1	2	3	4	5
		B	2	1	3	4	5
0	A	**0.00**	**1.00**	2.00	3.00	4.00	5.00
1	1	1.00	2.00	**1.00**	**2.00**	3.00	4.00
2	4	2.00	3.00	2.00	3.00	**2.00**	3.00
3	3	3.00	4.00	3.00	2.00	**3.00**	4.00
4	2	4.00	3.00	4.00	3.00	**4.00**	5.00
5	5	5.00	4.00	5.00	4.00	5.00	**4.00**

[...]

d)

D Matrix		0	1	2	3	4	5
		B	1	2	3	4	5
0	A	**0.00**	**1.00**	**2.00**	**3.00**	**4.00**	5.00
1	5	1.00	2.00	3.00	4.00	5.00	**4.00**
2	4	2.00	3.00	4.00	5.00	4.00	**5.00**
3	3	3.00	4.00	5.00	4.00	5.00	**6.00**
4	2	4.00	5.00	4.00	5.00	6.00	**7.00**
5	1	5.00	4.00	5.00	6.00	7.00	**8.00**

[...]

e)

D Matrix		0	1	2	3	4	5
		B	2	1	3	4	5
0	A	**0.00**	1.00	2.00	3.00	4.00	5.00
1	5	**1.00**	2.00	3.00	4.00	5.00	4.00
2	4	**2.00**	3.00	4.00	5.00	4.00	5.00
3	3	**3.00**	4.00	5.00	4.00	5.00	6.00
4	2	4.00	**3.00**	4.00	5.00	6.00	7.00
5	1	5.00	4.00	**3.00**	**4.00**	**5.00**	**6.00**

f)

D Matrix		0	1	2	3	4	5
		B	1	4	3	2	5
0	A	**0.00**	1.00	2.00	3.00	4.00	5.00
1	5	1.00	**2.00**	3.00	4.00	5.00	4.00
2	4	2.00	3.00	**2.00**	3.00	4.00	5.00
3	3	3.00	4.00	3.00	**2.00**	3.00	4.00
4	2	4.00	5.00	4.00	3.00	**2.00**	3.00
5	1	5.00	4.00	5.00	4.00	3.00	**4.00**

[...]

g)

D Matrix		0	1	2	3	4	5
		B	1	2	3	4	5
0	A	**0.00**	1.00	2.00	3.00	4.00	5.00
1	1	1.00	**0.00**	**1.00**	2.00	3.00	4.00
2	3	2.00	1.00	2.00	**1.00**	**2.00**	3.00
3	5	3.00	2.00	3.00	2.00	3.00	**2.00**
4	4	4.00	3.00	4.00	3.00	2.00	**3.00**
5	2	5.00	4.00	3.00	4.00	3.00	**4.00**

h)

D Matrix		0	1	2	3	4	5
		B	2	1	3	4	5
0	A	**0.00**	**1.00**	2.00	3.00	4.00	5.00
1	1	1.00	2.00	**1.00**	2.00	3.00	4.00
2	3	2.00	3.00	2.00	**1.00**	**2.00**	3.00
3	5	3.00	4.00	3.00	2.00	3.00	**2.00**
4	4	4.00	5.00	4.00	3.00	2.00	**3.00**
5	2	5.00	4.00	5.00	4.00	3.00	**4.00**

i)

D Matrix		0	1	2	3	4	5
		B	1	4	3	2	5
0	A	**0.00**	1.00	2.00	3.00	4.00	5.00
1	1	1.00	**0.00**	**1.00**	2.00	3.00	4.00
2	3	2.00	1.00	2.00	**1.00**	**2.00**	3.00
3	5	3.00	2.00	3.00	2.00	3.00	**2.00**
4	4	4.00	3.00	2.00	3.00	4.00	**3.00**
5	2	5.00	4.00	3.00	4.00	3.00	**4.00**

j)

D Matrix		0	1	2	3	4	5
		B	5	4	3	2	1
0	A	**0.00**	1.00	2.00	3.00	4.00	5.00
1	1	**1.00**	2.00	3.00	4.00	5.00	4.00
2	3	**2.00**	3.00	4.00	3.00	4.00	5.00
3	5	3.00	**2.00**	3.00	4.00	5.00	6.00
4	4	4.00	3.00	**2.00**	**3.00**	4.00	5.00
5	2	5.00	4.00	3.00	4.00	**3.00**	**4.00**

Im Folgenden wird das Vorgehen der oben dargestellten Matrizen näher beschrieben:

- *Matrix a)*

  In Sequenz A wird zuerst das Element „2“ gelöscht (dabei einstehen Kosten von „1“). Damit ist das Element „1“ in der Ausgangs- und der Zielsequenz identisch (es entstehen dabei also keine zusätzlichen Kosten). In einem nächsten Schritt wird das Element „2“ eingefügt (es entstehen Kosten von „1“, also insgesamt bis zu dieser Zelle von „2“). In den darauf folgenden Schritten entstehen keine weiteren Kosten, da die Werte „3“, „4“ und „5“ in Sequenz A und B identisch sind. Die beiden Sequenzen weisen also in Bezug auf diese drei Elemente ein gemeinsames Muster auf.

  Insgesamt sind zum Herstellen der Übereinstimmung („alignment“) von Sequenz A und Sequenz B zwei Operationen notwendig: *Löschen* und *Einfügen*. Dabei entstehen insgesamt Kosten von „2“.
- *Matrix b)*

  In Sequenz A und B befindet sich an erster Stelle das Element „1“, also ein identischer Wert. Eine Operation zur Transformation ist daher nicht notwendig. In einem zweiten Schritt wird in der Sequenz A das Element „4“ durch das Element „2“ ausgetauscht (Dabei entstehen Kosten für das Austauschen von „2“). Beim nächsten Schritt entsprechen sich das Elemente „3“ in Sequenz A und B, es ist daher keine Operationen notwendig und es entstehen keine weiteren Kosten. Im nächsten Schritt wird in der Sequenz A das Element „2“ durch das Element „4“ ausgetauscht (dabei entstehen Kosten für das Austauschen von „2“, also insgesamt bis zu dieser Zelle von „4“). Im letzten Schritt ist aufgrund des identischen Elementes „5“ keine Operation notwendig und es entstehen keine weiteren Kosten.

  Insgesamt sind zum Herstellen der Übereinstimmung („alignment“) von Sequenz A und Sequenz B zwei Substitutions-Operationen notwendig. Dabei entstehen insgesamt Kosten von „4“.
- *Matrix c)*

  In Sequenz A wird in einem ersten Schritt das Element „1“ eingefügt (dabei entstehen Kosten von „1“). Damit ist das Element „1“ in Sequenz A und Sequenz B identisch (es entstehen also keine weiteren Kosten). Im nächsten Schritt wird in Sequenz A das Element „3“ eingefügt (es entstehen dabei Kosten von „1“, also insgesamt bis zu dieser Zellen von „2“). Im folgenden Schritt entstehen keine Kosten, da das Element „4“ in Sequenz A und Sequenz B identisch ist. Im nächsten Schritt wird in Sequenz A das Element „3“ gelöscht (dabei entstehen Kosten von „1“, also insgesamt bis zu dieser Zelle von insgesamt „3“). Im folgenden Schritt wird in Sequenz A das Element „2“ gelöscht (dabei entstehen Kosten von „1“, also insgesamt bis zu dieser Zelle von „4“). Für den letzten Schritt entstehen keine weiteren Kosten, da das Element „5“ in beiden Sequenzen identisch ist.

  Insgesamt sind zum Herstellen der Übereinstimmung („alignment“) von Sequenz A und Sequenz B

vier Operationen notwendig: zweimal Einfügen und zweimal Löschen. Dabei entstehen insgesamt Kosten von „4".

- *Matrix d)*

  In Sequenz A werden in den ersten vier Schritten jeweils die Elemente „1", „2", „3" und „4" eingefügt (dabei entstehen jeweils Kosten von „1", also insgesamt von „4"). Im fünften Schritt entstehen keine Kosten, da das Element „5" in Sequenz A und Sequenz B identisch ist. In den nächsten vier Schritten werden in Sequenz A die Elemente „4", „3", „2" und „1" gelöscht (dabei entstehen jeweils Kosten von „1", also insgesamt von „4").

  Insgesamt sind zum Herstellen der Übereinstimmung („alignment") von Sequenz A und Sequenz B acht Operationen notwendig: viermal Einfügen und viermal Löschen. Dabei entstehen insgesamt Kosten von „8".[61]

- *Matrix e)*

  In Sequenz A werden in den ersten drei Schritten die Elemente „5", „4" und „3" gelöscht (dabei entstehen Kosten von jeweils „1", also insgesamt von „3"). Für die folgenden beiden Schritte entstehen keine weiteren Kosten, da die Elemente „2" und „1" in Sequenz A und Sequenz B identisch sind. In den nächsten drei Schritten werden in Sequenz A die Elemente „3", „4" und „5" eingefügt (dabei entstehen Kosten von jeweils „1", als insgesamt von „3").

  Insgesamt ist zum Herstellen der Übereinstimmung („alignment") von Sequenz A und Sequenz B sechs Operationen notwendig: dreimal Löschen und dreimal Einfügen. Dabei entstehen insgesamt Kosten von „6".

- *Matrix f)*

  In Sequenz A wird in einem ersten Schritt das Element „5" durch das Element „1" ausgetauscht (dabei entstehen Kosten für das Austauschen von „2"). In den folgenden drei Schritten entstehen keine weiteren Kosten, da die Elemente „4", „3" und „2" in Sequenz A und Sequenz B identisch sind. In einem letzten Schritt wird in Sequenz A das Element „1" durch das Element „5" substituiert (dabei entstehen Kosten für das Austauschen von „2").

  Insgesamt sind zum Herstellen der Übereinstimmung („alignment") von Sequenz A und Sequenz B zwei Operationen notwendig: zweimal Substitution. Dabei entstehen insgesamt Kosten von „4".

- *Matrix g)*

  In Sequenz A und B befindet sich an erster Stelle das Element „1", also ein identischer Wert. Eine Operation zur Transformation ist daher nicht notwendig. In einem zweiten Schritt wird in der Sequenz A das Element „2" eingefügt (dabei entstehen Kosten von „1"). Im nächsten Schritt entstehen keine weiteren Kosten, da das Element „3" in Sequenz A und Sequenz B identisch ist. Im nächsten Schritt wird in Sequenz A das Element „4" eingefügt (dabei entstehen Kosten von „1", also insge-

61 Diese Matrix entspricht der maximalen Distanz zwischen Sequenzen des Beispieldatensatzes, vgl. Ausgabedatei (output-file) Zeile 96: 63.

samt bis zu dieser Zelle von „2“). Im nächsten Schritt entstehen keine weiteren Kosten, das das Element „5“ in Sequenz A und Sequenz B identisch ist. In den letzten beiden Schritten wird in Sequenz A das Element „4“ und das Element „2“ gelöscht (dabei entstehen jeweils Kosten von „1“, also insgesamt von „2“).

Insgesamt sind zum Herstellen der Übereinstimmung („alignment“) von Sequenz A und Sequenz B vier Operationen notwendig: zweimal Einfügen und zweimal Löschen. Dabei entstehen insgesamt Kosten von „4“.

- *Matrix h)*

  In einem ersten Schritt wird in der Sequenz A das Element „2“ eingefügt (dabei entstehen Kosten von „1“). In den folgenden zwei Schritten entstehen keine weiteren Kosten, da das Element „1“ und das Element „3“ in Sequenz A und Sequenz B identisch sind. Im nächsten Schritt wird in Sequenz A das Element „4“ eingefügt (dabei entstehen Kosten von „1“, also insgesamt bis zu dieser Zelle von „2“). Im nächsten Schritt entstehen keine weiteren Kosten, da das Element „5“ in Sequenz A und Sequenz B identisch ist. In den letzten beiden Schritten werden in Sequenz A die Elemente „4“ und „2“ gelöscht (dabei entstehen jeweils Kosten von „1“, also insgesamt von „2“).

  Insgesamt sind zum Herstellen der Übereinstimmung („alignment“) von Sequenz A und Sequenz B vier Operationen notwendig: zweimal Einfügen und zweimal Löschen. Dabei entstehen insgesamt Kosten von „4“.

- *Matrix i)*

  In Sequenz A und B befindet sich an erster Stelle das Element „1“, also ein identischer Wert. Eine Operation zur Transformation ist daher nicht notwendig. In einem zweiten Schritt wird in der Sequenz A das Element „4“ eingefügt (dabei entstehen Kosten von „1“). Im nächsten Schritt entstehen keine weiteren Kosten, da das Element „3“ in Sequenz A und Sequenz B identisch sind. Im nächsten Schritt wird in Sequenz A das Element „2“ eingefügt (dabei entstehen Kosten von „1“, also insgesamt bis zu dieser Zelle von „2“). Im nächsten Schritt entstehen keine weiteren Kosten, das das Element „5“ in Sequenz A und Sequenz B identisch sind. In den letzten zwei Schritten werden in Sequenz A die Elemente „3“ und „4“ gelöscht (dabei entstehen Kosten von jeweils „1“, also insgesamt von „2“).

  Insgesamt sind zum Herstellen der Übereinstimmung („alignment“) von Sequenz A und Sequenz B vier Operationen notwendig: zweimal Einfügen und zweimal Löschen. Dabei entstehen insgesamt Kosten von „4“.

- *Matrix j)*

  In den ersten beiden Schritten werden in Sequenz A die Elemente „1“ und „3“ gelöscht (dabei entstehen Kosten von jeweils „1“, also insgesamt von „2“). In den nächsten zwei Schritten entstehen keine weiteren Kosten, da die Elemente „5“ und „4“ in der Sequenz A und Sequenz B identisch

sind. Im nächsten Schritt wird in Sequenz A das Element „3“ eingefügt (dabei entstehen Kosten von „1“; also insgesamt bis zu dieser Zelle von „3“). Im nächsten Schritt entstehen keine weiteren Kosten, da das Element „2“ in Sequenz A und Sequenz B identisch ist. Im letzten Schritt wird in der Sequenz A das Element „1“ eingefügt (dabei entstehen Kosten von „1“).
Insgesamt sind zum Herstellen der Übereinstimmung („alignment“) von Sequenz A und Sequenz B vier Operationen notwendig: zweimal Löschen und zweimal Einfügen. Dabei entstehen insgesamt Kosten von „4“.

Grundsätzlich ist an diesen Matrizen zur Ermittlung der Levenshtein-Distanzen zu erkennen, dass bei zugrunde legen der *default*-Substitutionskosten alle drei grundlegenden Operationen (*Einfügen*, *Löschen*, *Austausche*n) zum Erreichen eines „alignments“ der Sequenzen genutzt werden (vgl. Kap. 6.5: 53, *Relation Substitutionskosten - Indelkosten*).
Darüber hinaus ist erkennbar, dass einerseits unterschiedliche Wege benutzt werden, um diese Übereinstimmung zu erreichen und andererseits sich gleiche Gesamtkosten, d.h. gleiche Werte der Levenshtein-Distanz, aus unterschiedlichen Operationen zusammensetzen können. So errechnet sich beispielsweise die Levenshtein-Distanz von „4“ in Matrix b) durch zweimalige Substitution, in Matrix c) durch zweimaliges Einfügen und zweimaliges Löschen.
Im Gegensatz zur Veranschaulichung der Ermittlung der Levenshtein-Distanz in Tabelle 9: 46 werden in der obigen Darstellung der Matrizen nicht die Werte für Indel bzw. Substitution in die Ecken der Zellen eingetragen, sondern nur die bis zu dieser Zelle entstehenden Kosten.
Während die Matrizen a – j die Ermittlung der Levenshtein-Distanz für den paarweisen Sequenzvergleich dokumentieren, wird das Ergebnis dieses Ermittlungsprozesses in die *Levenshtein*-Distanzmatrix eingetragen. Diese auf der Grundlage des Optimal-Matching Verfahrens errechneten Levenshtein-Distanzen werden von TDA in Form einer *Tabelle* ausgegeben (vgl. Tabelle 3). Dabei handelt es sich lediglich um eine andere Form der Darstellung; die Distanztabelle und die Distanzmatrix enthalten die gleichen Informationen. Bei der geringen Anzahl von 11 zu vergleichenden Sequenzen ist eine Distanzmatrix noch darstellbar (vgl. Tabelle S. 73), diese Darstellbarkeit stößt jedoch bei einer größeren Anzahl von Sequenzen an Grenzen, wenn z.B. einige hundert Sequenzen miteinander verglichen werden. In diesem Fall würde die Distanzmatrix aus einigen hundert Spalten und Zeilen bestehen.
In der ersten Spalte der Distanztabelle befindet sich die Fallnummer der Ausgangssequenz, in der zweiten Spalte die Fallnummer der Zielsequenz, in der dritten Spalte befindet sich die Anzahl der Elemente der Ausgangssequenz, in der vierten Spalte befindet sich die Anzahl der Elemente der Zielsequenz, in der fünften Spalte befindet sich die Levenshtein-Distanz als Maßzahl der Distanz zwischen Ausgangs- und Zielsequenz. Eine Distanzmatrix wird von TDA nicht ausgegeben, sie kann jedoch bei der vorliegenden geringen Anzahl

von Sequenzen der Beispieldaten auf Grundlage der obigen Tabelle konstruiert werden und wird hier aus Gründen der Verdeutlichung dargestellt (vgl. Tab. 4).

2	1	5	5	0.00
3	1	5	5	0.00
3	2	5	5	0.00
4	1	5	5	2.00
4	2	5	5	2.00
4	3	5	5	2.00
5	1	5	5	2.00
5	2	5	5	2.00
5	3	5	5	2.00
5	4	5	5	0.00
6	1	5	5	2.00
6	2	5	5	2.00
6	3	5	5	2.00
6	4	5	5	0.00
6	5	5	5	0.00
7	1	5	5	4.00
7	2	5	5	4.00
7	3	5	5	4.00
7	4	5	5	4.00
7	5	5	5	4.00
7	6	5	5	4.00
8	1	5	5	4.00
8	2	5	5	4.00
8	3	5	5	4.00
8	4	5	5	4.00
8	5	5	5	4.00
8	6	5	5	4.00
8	7	5	5	0.00
9	1	5	5	4.00
9	2	5	5	4.00
9	3	5	5	4.00
9	4	5	5	4.00
9	5	5	5	4.00
9	6	5	5	4.00
9	7	5	5	0.00
9	8	5	5	0.00
10	1	5	5	8.00
10	2	5	5	8.00
10	3	5	5	8.00
10	4	5	5	6.00

10	5	5	5	6.00
10	6	5	5	6.00
10	7	5	5	4.00
10	8	5	5	4.00
10	9	5	5	4.00
11	1	5	5	4.00
11	2	5	5	4.00
11	3	5	5	4.00
11	4	5	5	4.00
11	5	5	5	4.00
11	6	5	5	4.00
11	7	5	5	4.00
11	8	5	5	4.00
11	9	5	5	4.00
11	10	5	5	4.00

Tabelle 3: Levenshtein Distanzmatrix (default-Substitutionskosten, in tabellarischer Form)

Diese Distanzmatrix (vgl. Tab. 4) stellt das Ergebnis der Optimal-Matching Analyse dar: in ihr wird die Distanz von Sequenzen durch die Maßzahl der Levenshtein-Distanz dargestellt.

	1	2	3	4	5	6	7	8	9	10	11
1	-										
2	0.00	-									
3	0.00	0.00	-								
4	2.00	2.00	2.00	-							
5	2.00	2.00	2.00	0.00	-						
6	2.00	2.00	2.00	0.00	0.00	-					
7	4.00	4.00	4.00	4.00	4.00	4.00	-				
8	4.00	4.00	4.00	4.00	4.00	4.00	0.00	-			
9	4.00	4.00	4.00	4.00	4.00	4.00	0.00	0.00	-		
10	8.00	8.00	8.00	6.00	6.00	6.00	4.00	4.00	4.00	-	
11	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	-

Tabelle 4: Levenshtein Distanzmatrix (default-Substitutionskosten)

Durch das in diesem Kapitel dargestellte Verfahren der Optimal-Matching Analyse werden also ausgehend von den grundlegenden Operationen *Einfügen*, *Löschen* und *Austauschen* Sequenzen paarweise verglichen und deren Distanz durch die Maßzahl der Levenshtein-Distanz ausgedrückt. Damit wird die eingangs dieses Kapitels formulierte Frage beantwortet, auf welche Weise Sequenzen verglichen werden und wie deren Unähnlichkeit festgestellt wird (vgl. Kap. 6: 41).

Abschließend wird nun auf Grundlage der errechneten Distanzmatrix eine Clusteranalyse nach dem Ward-Verfahren durchgeführt. Die nachstehende Abbildung 12:*Dendogramm (513) default-Kosten* verdeutlicht den Agglomerationsprozess anhand eines Dendogramms.

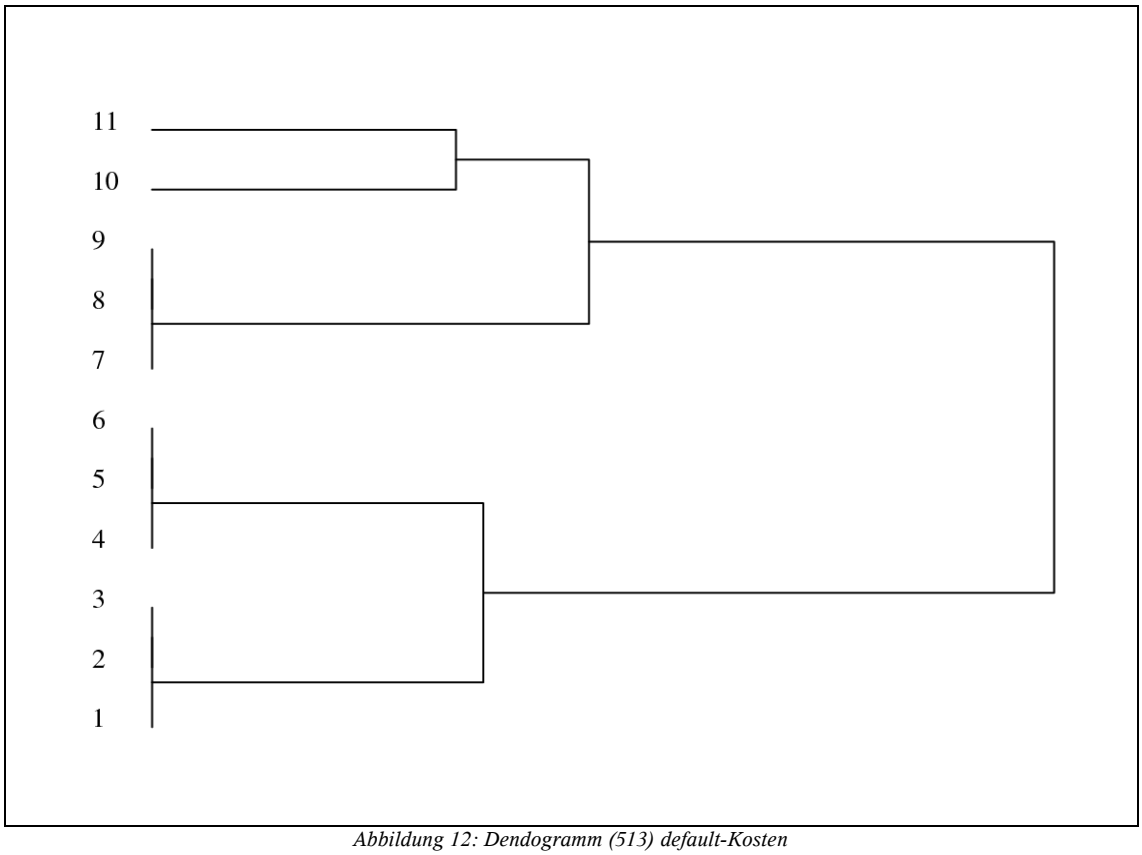

Abbildung 12: Dendogramm (513) default-Kosten

Grundsätzlich ist an diesem Dendogramm ablesbar, dass im Rahmen der Clusteranalyse zunächst diejenigen Sequenzen einem einheitlichen Cluster zugeordnet werden, die aus identischen Abfolgen von Elementen bestehen, und deren Distanz daher „0“ beträgt (Fälle 1, 2, 3; Fälle 4, 5, 6; Fälle 7,8,9). Im nächsten Schritt wird der Fall 11 und 10 fusioniert; dann die Cluster 1-2-3 und 4-5-6, daran anschließend die Cluster 7-8-9 mit 10-11, und in einem letzten Schritt alle Cluster zu einem übergeordneten Cluster.

7.2 Datenbasierte Substitutionskosten

Analog zur Darstellung der Ermittlung der Levenshtein-Distanz aufgrund der *default*-Substitutionskosten wird in diesem Abschnitt die Ermittlung der Levenshtein-Distanz aufgrund der datenbasierten Substitutionskosten dargestellt.

Die vollständige Ausgabedatei (*.tst) befindet sich im Anhang (s. Kapitel 17.8: 233). Sie unterscheidet sich von der Ausgabedatei auf Grundlage der *default*-Substitutionskosten lediglich durch die Dokumentation der Verwendung der datenbasierten Substitutionskosten in Zeile 86: 234 („Substitution cost based on data, type 2."), sowie durch die in Zeile 95: 234 dokumentierte maximale Distanz zwischen Sequenzen („Maximum distance between sequences 10 and 1: 6.90909"). Das allgemeine Vorgehen der Optimal-Matching Analyse ist mit Ausnahme der verwendeten Substitutionskosten identisch, Unterschiede bestehen jedoch im Ergebnis.

Analog zur Darstellung der Ausgabedatei (*.tst) für die *default*-Substitutionskosten wird im Folgenden die Ausgabedatei für die datenbasierten Kosten dargestellt. In ihr werden die datenbasierten Substitutionskosten in Form einer Matrix dokumentiert (Substitutionskosten-Matrix), sowie die Matrizen zur Errechnung der Levenshtein-Distanz für den paarweisen Sequenzvergleich auf Grundlage dieser Substitutionsmatrix.

```
Optimal matching test output file.
Number of states: 5
Max sequence lenght: 5

Indel cost
1 1 1 1 1

Substitution cost
0 1.3 1.6 1.7 2
1.3 0 1.33636 1.90909 1.7
1.6 1.33636 0 1.09091 1.90909
1.7 1.90909 1.09091 0 0.454545
2 1.7 1.90909 0.454545 0


[...]

a)
D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       2       3       4       5
          ------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    2 |   1.00    1.30    1.00    2.00    3.00    4.00
   2    1 |   2.00    1.00    2.00    2.60    3.60    4.60
   3    3 |   3.00    2.00    2.34    2.00    3.00    4.00
   4    4 |   4.00    3.00    3.34    3.00    2.00    3.00
   5    5 |   5.00    4.00    4.34    4.00    3.00    2.00

[...]


b)
```

```
D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       2       3       4       5
          ------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    1 |   1.00    0.00    1.00    2.00    3.00    4.00
   2    4 |   2.00    1.00    1.91    2.09    2.00    3.00
   3    3 |   3.00    2.00    2.34    1.91    2.91    3.91
   4    2 |   4.00    3.00    2.00    2.91    3.82    4.61
   5    5 |   5.00    4.00    3.00    3.91    3.36    3.82

[...]
```

c)

```
D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       2       1       3       4       5
          ------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    1 |   1.00    1.30    1.00    2.00    3.00    4.00
   2    4 |   2.00    2.30    2.00    2.09    2.00    3.00
   3    3 |   3.00    3.30    3.00    2.00    3.00    3.91
   4    2 |   4.00    3.00    4.00    3.00    3.91    4.70
   5    5 |   5.00    4.00    5.00    4.00    3.45    3.91

[...]
```

d)

```
D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       2       3       4       5
          ------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    5 |   1.00    2.00    2.70    3.70    3.45    4.00
   2    4 |   2.00    2.70    3.70    3.79    3.70    3.91
   3    3 |   3.00    3.60    4.04    3.70    4.70    4.91
   4    2 |   4.00    4.30    3.60    4.60    5.60    5.91
   5    1 |   5.00    4.00    4.60    5.20    6.20    6.91

[...]
```

e)

```
D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       2       1       3       4       5
          ------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    5 |   1.00    1.70    2.70    3.70    3.45    4.00
   2    4 |   2.00    2.70    3.40    3.79    3.70    3.91
   3    3 |   3.00    3.34    4.30    3.40    4.40    4.91
   4    2 |   4.00    3.00    4.00    4.40    5.31    5.91
```

```
   5   1 |   5.00    4.00    3.00    4.00    5.00    6.00

[...]

f)
D Matrix        0       1       2       3       4       5
         ------------------------------------------------
                B       1       4       3       2       5
         ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   5 |   1.00    2.00    1.45    2.45    3.45    4.00
   2   4 |   2.00    2.70    2.00    2.55    3.55    3.91
   3   3 |   3.00    3.60    3.00    2.00    3.00    4.00
   4   2 |   4.00    4.30    4.00    3.00    2.00    3.00
   5   1 |   5.00    4.00    5.00    4.00    3.00    4.00

[....]

g)
D Matrix        0       1       2       3       4       5
         ------------------------------------------------
                B       1       2       3       4       5
         ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    0.00    1.00    2.00    3.00    4.00
   2   3 |   2.00    1.00    1.34    1.00    2.00    3.00
   3   5 |   3.00    2.00    2.34    2.00    1.45    2.00
   4   4 |   4.00    3.00    3.34    3.00    2.00    1.91
   5   2 |   5.00    4.00    3.00    4.00    3.00    2.91

[...]

h)
D Matrix        0       1       2       3       4       5
         ------------------------------------------------
                B       2       1       3       4       5
         ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    1.30    1.00    2.00    3.00    4.00
   2   3 |   2.00    2.30    2.00    1.00    2.00    3.00
   3   5 |   3.00    3.30    3.00    2.00    1.45    2.00
   4   4 |   4.00    4.30    4.00    3.00    2.00    1.91
   5   2 |   5.00    4.00    5.00    4.00    3.00    2.91

[...]

i)
```

```
D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       4       3       2       5
          ------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    1 |   1.00    0.00    1.00    2.00    3.00    4.00
   2    3 |   2.00    1.00    1.09    1.00    2.00    3.00
   3    5 |   3.00    2.00    1.45    2.00    2.70    2.00
   4    4 |   4.00    3.00    2.00    2.55    3.55    3.00
   5    2 |   5.00    4.00    3.00    3.34    2.55    3.55

[...]

j)
D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       5       4       3       2       1
          ------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    1 |   1.00    2.00    2.70    3.60    4.30    4.00
   2    3 |   2.00    2.91    3.09    2.70    3.70    4.70
   3    5 |   3.00    2.00    3.00    3.70    4.40    5.40
   4    4 |   4.00    3.00    2.00    3.00    4.00    5.00
   5    2 |   5.00    4.00    3.00    3.34    3.00    4.00
```

Da die grundlegende Interpretation der Ermittlung der Levenshtein-Distanz dieser datenbasierten Matrizen denen der Matrizen auf Grundlage der default-Substitutionskosten entspricht, sind in den obigen Matrizen jeweils lediglich die Wege zur Ermittlung der Levenshtein-Distanz durch Fettformatierung gekennzeichnet. Auf eine ausführliche Beschreibung jeder einzelnen Matrix wird an dieser Stelle verzichtet.

Neben den einzelnen Matrizen zur Ermittlung der Levenshtein-Distanz werden in dieser Datei die datenbasierten Substitutionskosten dokumentiert: Da die Substitutionskosten-Matrix in der Ausgabedatei unformatiert ausgegeben wird (die Zellen sind durch ein Leerzeichen getrennt) wird hier zur besseren Lesbarkeit eine formatierte Darstellung eingefügt:

Sequenz-Nr.	1	2	3	4	5
1	0	1.3	1.6	1.7	2
2	1.3	0	1.33636	1.90909	1.7
3	1.6	1.33636	0	1.09091	1.90909
4	1.7	1.90909	1.09091	0	0.454545
5	2	1.7	1.90909	0.454545	0

Während bei den default-Substitutionskosten die Tabelle für jede Substitution den Wert „2“ enthält (vgl. S. 65), enthält die Tabelle der datenbasierten Substitutionskosten die konkreten Werte in Abhängigkeit der Übergangswahrscheinlichkeiten der Zustandswechsel (vgl. 6.4: 50). Niedrige Substitutionskosten verweisen dabei auf häufig vorkommende Übergänge in den empirischen Daten; hohe Substitutionskosten verweisen auf weniger häufig vorkommende Übergänge.

- Der Übergang von Element „1“ zu Element „5“ (und umgekehrt) ist im Beispieldatensatz nicht vorhanden. Daher beträgt der Wert für deren Substitution „2“ (dies entspricht den maximalen Substitutionskosten).
- Der Übergang von Element „4“ zu Element „5“ (um umgekehrt) wird auf Grundlage der empirischen Übergangswahrscheinlichkeiten der Substitution dieser Elemente der Wert „0.454545“ zugewiesen (dies entspricht den niedrigsten Substitutionskosten für diese Beispielsequenzen). Dabei ist der Übergang „4“ zu „5“ genau sechs Mal im Datensatz enthalten (je einmal in den Sequenzen 1, 2, 3, 4, 5, 6) und der Übergang „5“ zu „4“ genau zwei mal (je einmal in den Sequenzen 10 und 11). Dies entspricht insgesamt 8 von 44 in den Beispielsequenzen enthaltenen Übergängen.

Die auf datenbasierten Substitutionskosten ermittelten Werte der Levenshtein-Distanz werden in folgender Tabelle dokumentiert.

2	1	5	5	0.00
3	1	5	5	0.00
3	2	5	5	0.00
4	1	5	5	2.00
4	2	5	5	2.00
4	3	5	5	2.00
5	1	5	5	2.00
5	2	5	5	2.00
5	3	5	5	2.00
5	4	5	5	0.00
6	1	5	5	2.00
6	2	5	5	2.00
6	3	5	5	2.00
6	4	5	5	0.00
6	5	5	5	0.00
7	1	5	5	3.82
7	2	5	5	3.82
7	3	5	5	3.82
7	4	5	5	3.91
7	5	5	5	3.91
7	6	5	5	3.91
8	1	5	5	3.82

8	2	5	5	3.82
8	3	5	5	3.82
8	4	5	5	3.91
8	5	5	5	3.91
8	6	5	5	3.91
8	7	5	5	0.00
9	1	5	5	3.82
9	2	5	5	3.82
9	3	5	5	3.82
9	4	5	5	3.91
9	5	5	5	3.91
9	6	5	5	3.91
9	7	5	5	0.00
9	8	5	5	0.00
10	1	5	5	6.91
10	2	5	5	6.91
10	3	5	5	6.91
10	4	5	5	6.00
10	5	5	5	6.00
10	6	5	5	6.00
10	7	5	5	4.00
10	8	5	5	4.00
10	9	5	5	4.00
11	1	5	5	2.91
11	2	5	5	2.91
11	3	5	5	2.91
11	4	5	5	2.91
11	5	5	5	2.91
11	6	5	5	2.91
11	7	5	5	3.55
11	8	5	5	3.55
11	9	5	5	3.55
11	10	5	5	4.00

Tabelle 5: Levenshtein Distanz (datenbasiert, in tabellarischer Form)

Transformiert man die obige Tabelle 5 analog dem Vorgehen im vorigen Kapitel in die Form einer Matrix, erhält man folgende Matrix-Darstellung (vgl. Tab. 81):

	1	2	3	4	5	6	7	8	9	10	11
1	-										
2	0.00	-									
3	0.00	0.00	-								
4	2.00	2.00	2.00	-							
5	2.00	2.00	2.00	0.00	-						
6	2.00	2.00	2.00	0.00	0.00	-					
7	3.82	3.82	3.82	3.91	3.91	3.91	-				
8	3.82	3.82	3.82	3.91	3.91	3.91	0.00	-			
9	3.82	3.82	3.82	3.91	3.91	3.91	0.00	0.00	-		
10	6.91	6.91	6.91	6.00	6.00	6.00	4.00	4.00	4.00	-	
11	2.91	2.91	2.91	2.91	2.91	2.91	3.55	3.55	3.55	4.00	-

Tabelle 6: Levenshtein Distanzmatrix (datenbasierte Substitutionskosten)

Bildet die oben abgebildete Tabelle 6 den Ausgangspunkt für eine Clusteranalyse nach dem Ward-Verfahren, kann der Agglomerationsprozess anhand des folgenden Dendogramms verdeutlicht werden.

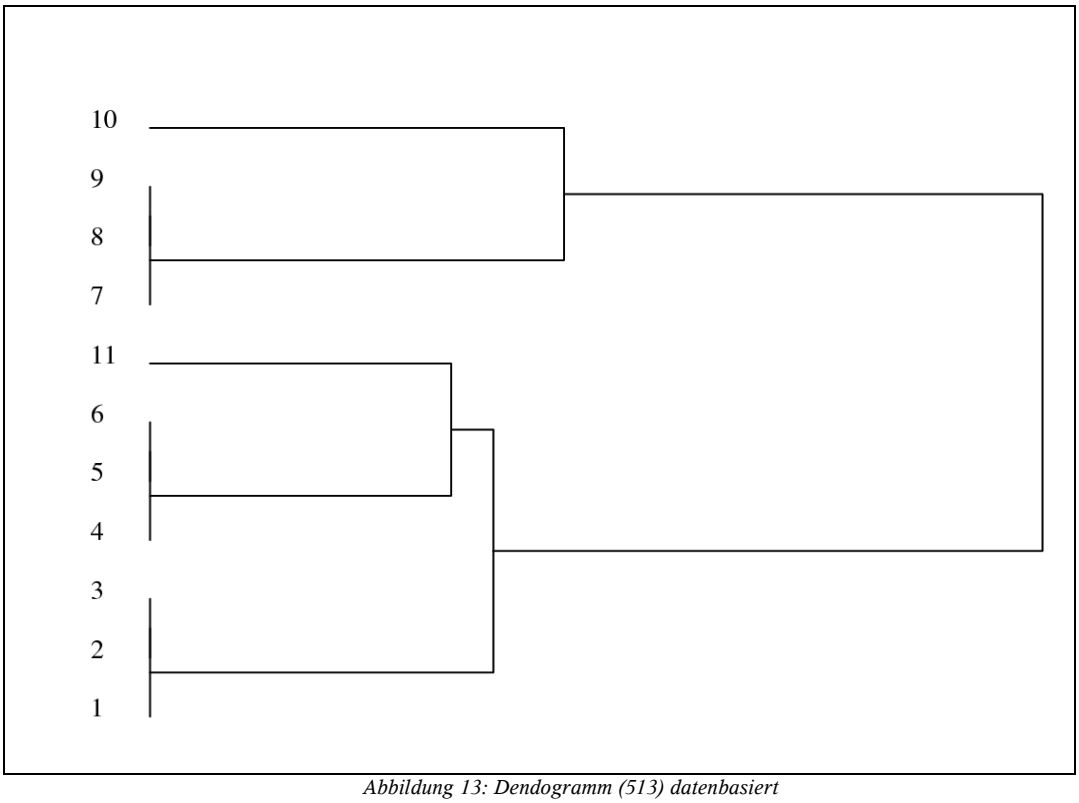

Abbildung 13: Dendogramm (513) datenbasiert

7.3 Substitutionskosten als absolute Differenz

Auf eine ausführliche Darstellung der Ermittlung der Levenshtein-Distanz auf Grundlage der Substitutionskosten als absoluter Differenz wird an dieser Stelle verzichtet.[62] Da die Distanzmatrix auf Grundlage der *default*-Substitutionskosten für den zu Grunde liegenden Datensatz identisch ist mit der Distanzmatrix auf Grundlage der Substitutionskosten als *absoluter Differenz*, wird die letztere bei der weiteren Analyse nicht weiter berücksichtigt. Das genaue Vorgehen zur Ermittlung der Levenshtein-Distanz ist darüber hinaus für diesen Fall der *absoluten Differenz* nicht transparent, das in der Ausgabedatei von TDA (*.tst) im Gegensatz zum Vorgehen für den Fall der *default*- bzw. der *datenbasierten* Substitutionskosten weder eine Dokumentation der Substitutionsmatrix noch die konkreten Matrizen zur Ermittlung der Levenshtein-Distanz enthalten ist.

Die vollständige Ausgabedatei (*.tst) für die Verwendung von Substitutionskosten als absoluter Differenz befindet sich im Anhang 17.9: 235, die Tabelle mit den Levenshtein-Distanzen in Anhang in Kapitel 17.10: 237.

7.4 Interpretation und Fazit

In diesem Kapitel wurde die Funktionsweise des Optimal-Matching Algorithmus am Beispiel des Effektes unterschiedlicher Definition der Substitutionskosten auf das Ergebnis der Clusteranalyse beispielhaft verdeutlicht. Dabei ist der Einfluss der Definition der Substitutionskosten auf das konkrete Vorgehen zum Erreichen des „alignments“ sowie auf den Wert der Levenshtein-Distanz deutlich geworden. Die Ergebnisse der *default*-Substitutionskosten sowie der *datenbasierten* Substitutionskosten sind in Tabelle 7: *Distanzen im Vergleich: default- und datenbasierte Substitutionskosten* gegenüber gestellt.

Es wird deutlich, dass die auf datenbasierten Substitutionskosten errechneten Distanzen *differenzierter* in sofern sind, als dass eine größere Anzahl unterschiedlicher Werte auftritt (vier unterschiedliche Werte bei den datenbasierten im Vergleich zu acht unterschiedlichen Werten bei den *default*-Definition). Darüber hinaus wird deutlich, dass die errechneten Distanzen nicht vollkommen unterschiedlich sind, sondern die datenbasierte Werte für vier Sequenzen identisch und für vier Sequenzen leicht unter den Werten der *default*-Definition liegen. Die maximale Differenz für die Distanzen im Beispieldatensatz liegt bei 1.09 für die Distanzen der Matrizen d, g und h.

62 In der TDA Syntax wird die Berechnung der Substitutionskosten als absoluter Differenz innerhalb des *seqm*-Befehls durch den Parameter *scost=1* definiert.

Anhand der ermittelten Distanzen für die paarweisen Sequenzvergleiche auf Grundlage der *default*-Substitutionskosten und datenbasierten Substitutionskosten wird deutlich, dass eine Übereinstimmung der Distanzen für die Fälle besteht, in denen die Angleichung der Sequenz A und der Sequenz B ausschließlich durch die Operationen *Einfügen* und *Löschen* erreicht wird. Für den Fall der *default*- als auch der datenbasierten Substitutionskosten beträgt der Wert für die Indelkosten stets „1“ (dies ist z.B. in Matrix a der Fall).

Matrix	Distanz (*default*)[63]	Distanz (datenbasiert)[64]
a)	2	2
b)	4	3.82
c)	4	3.91
d)	8	6.91
e)	6	6
f)	4	4
g)	4	2.91
h)	4	2.91
i)	4	3.55
j)	4	4

Tabelle 7: Distanzen im Vergleich: default- und datenbasierte Substitutionskosten

Dies gilt jedoch nur unter der Voraussetzung, dass die Verwendung der Operation Indel kostengünstiger ist als die Verwendung der Operation Substitution. An der Matrix d ist genau das Gegenteil erkennbar und damit wird an dieser Stelle der Einfluss der unterschiedlichen Definition der Substitutionskosten auf den Weg zur Ermittlung der Levenshtein-Distanz besonders deutlich.

- Bei dem Vorgehen auf Grundlage der *default-Substitutionskosten* werden zunächst in Sequenz A die Elemente „1“, „2“, „3“, „4“ eingefügt, das Element „5“ ist damit in beiden Sequenzen identisch, abschließend werden in Sequenz A die Elemente „4“, „3“, „2“, „1“ gelöscht, um eine Übereinstimmung („alignment“) von Sequenz A und Sequenz B zu erreichen.
- Bei dem Vorgehen auf Grundlage der *datenbasierten Substitutionskosten* werden zunächst in Sequenz die Elemente „1“, „2“, „3“ eingefügt. Von dieser Zelle an unterscheiden sich die Vorgehensweisen: die Elemente „5“ und „4“ in Sequenz A werden ausgetauscht durch die Elemente „4“ und „5“. Die Kosten für die Substitution dieser Elemente betragen je „0.454545“, so dass dieses Vorgehen gegenüber dem Löschen und Einfügen der betreffenden Elemente kostengünstiger ist. Abschließend werden die Elemente „3“, „2“, „1“ gelöscht. Damit ist die *Substitution* von Elementen in diesem Fall mit weniger Kosten verbunden als das *Löschen* oder *Einfügen* von Elementen.

63 Distanzen auf Grundlage der default-Substitutionskosten (vgl. Kap. 7.1: 62f.).

64 Distanzen auf Grundlage der datenbasierten Substitutionskosten (vgl. Kap.7.2: 74f.).

Im Gegensatz zum Beispiel Matrix d hat in Matrix b die Definition der Substitutionskosten keinen Einfluss auf das konkrete Vorgehen zur Ermittlung der Levenshtein-Distanz. Die unterschiedlichen Levenshtein-Distanzen für den Fall der *default*- und der datenbasierten Substitutionskosten sind auf die unterschiedlichen Substitutionskosten zurückzuführen, nicht auf das konkrete Vorgehen als solches: es werden jeweils zwei Substitutionen verwendet, um eine Übereinstimmung der Sequenz A und der Sequenz B zu erreichen. Diese zwei Substitutionen entsprechen im Fall der *default*-Substitutionskosten insgesamt dem Wert „4“ (jeweils „2“ pro Substitution), im Fall der datenbasierten Substitutionskosten für das Austauschen der Elemente „2“ und „4“ insgesamt dem Wert „3.81“ (jeweils „1.91“ pro Substitution).

An dieser Stelle ist besonders auf Matrix f hinzuweisen, da der datenbasiert ermittelte Wert für die Substitution der Elemente „1“ und „5“ mit dem Wert „2“ exakt den default-Substitutionskosten entspricht. Daher bestehen in diesem Fall weder im konkreten Vorgehen zur Ermittlung der Levenshtein-Distanz noch in den Kosten der betreffenden Substitution Unterschiede.

Der Effekt der unterschiedlichen Definition von Substitutionskosten in Form unterschiedlicher Levenshtein-Distanzen zeigt sich besonders am Ergebnis der Clusterlösung der Sequenzen 10 und 11. Diese werden aufgrund der unterschiedlichen Levenshtein-Distanz in Abhängigkeit der Kostendefinitionen unterschiedlich fusioniert:

- Auf Grundlage der *default*-Definition wird zunächst die Sequenz 10 mit der Sequenz 11 zu einem Cluster fusioniert. Danach werden die Sequenzen 1-2-3 mit den Sequenzen 4-5-6 fusioniert. Daran anschließend werden die Sequenzen 10-11 mit den Sequenzen 7-8-9 fusioniert (vgl. 12: 74).
- Auf Grundlage der datenbasierten Definition wird die Sequenz 11 mit zunächst mit den Sequenzen 4-5-6 fusioniert, danach dieses neu gebildete Cluster mit den Sequenzen 1-2-3. Anschließend wird die Sequenz 10mit den Sequenzen 7-8-9 fusioniert (vgl. Abb. 13: 81)

An der Clusterlösung der Sequenzen 10 und 11 wird zudem ein grundlegendes Kennzeichen des Ward-Clusterverfahrens deutlich: Die Fusionierung zweier Cluster ist abhängig von der Zunahme der Fehlerquadratsumme des neu fusionierten Clusters (vgl. Kap. 9.2; *Clusteranalyse im Rahmen der Navigationsanalyse*).

Zum Abschluss dieses Kapitels ist nochmals darauf hinzuweisen, dass es sich bei den in diesem Kapitel analysierten Sequenzen um eine beispielhafte Demonstration des Vorgehens des Optimal-Matching Algorithmus anhand eines exemplarischen, fiktiven Datensatzes handelt. Die Analyse empirischer Navigationssequenzen als Form von Verhaltensspuren wird in den folgenden Kapitel dargestellt: Die Definition der Substitutionskosten für die Daten der Navigationsanalyse wird in Kapitel 9.1: 96 erläutert, die Wahl des Clusterverfahrens in Kapitel 9.3.5: 120 sowie die Wahl der verwendeten Fusionsebene in Kapitel 9.4: 123.

8 Ereignisdatenanalyse

In diesem Kapitel wird mit der Ereignisdatenanalyse (vgl. Blossfeld / Rohwer 2002)[65] ein spezifisches Verfahren zur Analyse von Verlaufsdaten dargestellt. Einerseits wird damit die Sequenzdatenanalyse auf Grundlage der Methode des Optimal-Matching in einem erweiterten forschungsmethodologischen Kontext verortet, andererseits wird durch diese Einordnung und Abgrenzung eine erweiterte Perspektive auf die Methode des Optimal-Matching ermöglicht.

Auf weitere in den Sozialwissenschaften verwendete Methoden zur Analyse zeitlicher Verläufe, z.B. in Form von Regressions- bzw. Pfadanalysen (Dependenzanalysen)[66] wird in diesem Kapitel nicht näher eingegangen. Pfadanalyse und Ereignisdatenanalyse beruhen zwar auf gemeinsamen methodologischen Grundlagen, unterschieden sich jedoch hinsichtlich der Möglichkeit der Berücksichtigung der zeitlichen Struktur des Verlaufs, sowie der Verweildauer in den einzelnen Zuständen: Im Gegensatz zu deren zentraler Bedeutung in der Ereignisdatenanalyse finden zeitliche Faktoren in Regressions- und Pfadanalysen grundsätzlich keine Berücksichtigung (vgl. Brüderl / Scherer 2005; Backhaus / Erichson / Plinke / Wulff 2000).

Bei allen methodologischen Unterschieden der Ereignis- und der Sequenzdatenanalyse ist jedoch deren *komplementärer*, sich ergänzender Charakter zu betonen (vgl. Sackmann / Wingens 2001), beispielsweise im Rahmen triangulativer Forschungsansätze. Ein besonders gelungenes Beispiel der Kombination ereignis- und sequenzdatenanalytischer Vorgehensweisen im Bereich der Lebenslaufforschung stellen die Untersuchungen von Windzio (2001) dar.

Im sozialwissenschaftlichen Bereich und dort vor allem im Bereich der Soziologie nimmt die Analyse von *verlausfsbezogenen Daten* seit den 1980er Jahren stetig zu, vor allem in der *soziologischen Lebenslaufforschung* auf Grundlage der Ereignisdatenanalyse (vgl. Blossfeld / Rohwer 2002; Sackmann / Wingens 2001; Brüderl / Scherer 2005). Mit der Fokussierung der quantitativen Analysen von Ereignissen des Lebenslaufs grenzt sie sich methodisch vor allem von der qualitativ orientierten Biographieforschung ab.

Allgemein besteht das methodologische Vorgehen der Ereignisdatenanalyse in der Analyse von Faktoren, die die Wahrscheinlichkeit des Auftretens von spezifischen Ereignissen bedingen: „Welche Einflüsse führen dazu, daß ein Individuum eine bestimmte berufliche Stellung erreicht? Welche Ausprägungen müssen be-

65 Die Begriffe „Ereignisanalyse“ und „Ereignisdatenanalyse“ werden im Folgenden synonym verwendet als Übersetzung des Begriffes „Event history analysis“, bei dem im Gegensatz zu den deutschen Übersetzungen der Bezug der Ereignisse zu einer grundlegenden Zeitachse deutlicher zum Ausdruck kommt.

66 Pfadanalysen (Dependenzanalysen) sind auf Regressions- und Korrelationsanalysen beruhende Modelle zur Analyse und Beschreibung von Abhängigkeitsbeziehungen in einem Set von Variablen. Dabei können direkte und indirekte Wirkungen der Variablen analysiert werden. Verwendung findet die Pfadanalyse in den Sozialwissenschaften vor allem bei der Konstruktion von Kausalmodellen (vgl. Bortz 1999, Backhaus et al. 2000).

stimmte Faktoren aufweisen, damit diese berufliche Stellung erreicht wird?“ (Aisenbrey 2001: 111). Typische *Forschungsfragen* beziehen sich auf den Übergang von beruflicher Ausbildung in die Erwerbstätigkeit (Mowitz-Lambert 2001), auf den Einfluss von Arbeitslosigkeit auf den weiteren Erwerbsverlauf (Windzio 2001), auf Armutssequenzen im Lebenslauf am Beispiel des Verlassens der Sozialhilfe (Hagen / Niemann 2001) oder auf kriminelle Phasen im Lebenslauf (Böttger 2001).

Den zentralen Forschungsansatz der Ereignisanalyse fassen Blossfeld / Rohwer (2002: 38; Hervorhebung im Original) wie folgt zusammen: „Event history analysis studies *transitions* across a set of discrete states, including the length of *time intervals* between entry to and exit from specific states. The basic analytical framework is a state space and a time axis.“

Grundlegendes Kennzeichen dieses Vorgehens ist die Analyse von *Übergängen* zwischen definierten Zuständen, wobei sich die Übergänge auf das *Verlassen* eines *Ausgangszustandes* („origin state“) und das *Eintreten* in einen *Zielzustand* („destination state“) beziehen. Diese Zustände sind *diskret* und Teil eines definierten Raums möglicher Zustand (*Zustandsraum*).

Bei der Analyse von Übergängen wird der *Faktor Zeit* auf zweierlei Weise berücksichtigt: Als Zeitspanne, die eine Untersuchunsgseinheit vom Eintritt bis zum Austritt in einem spezifischen Zustand verbringt („*episode*“), sowie als grundlegende *Zeitachse* des Verlaufs mit der Möglichkeit der Analyse der Reihenfolge des zeitlichen Auftretens von Übergängen und Ereignissen.

Die *kleinste Analyseeinheit* der Ereignisanalyse stellt ein Prozess dar, der lediglich aus *einer* Episode und *zwei* Zuständen besteht, und genauer: dem Eintritt in einen Ausgangszustand, der Verweildauer in diesem Zustand und dem Übergang in einen Zielzustand.

Sind Übergänge zu mehr als einem Zielzustand möglich, spricht man von „multistate-models“ oder „models with competing events or risks“ (Blossfeld / Rohwer 2002: 39). Bewegt sich die Untersuchungseinheit wiederholt zwischen unterschiedlichen Zuständen, wird dies als „multistate-multi-episode process“ (Blossfeld / Rohwer 2002: 39) bezeichnet.

In dieser Perspektive der Ereignisdatenanalyse kann die Mikronavigation innerhalb der Navigationsanalyse als ein solcher *multistate-multi-episode* Prozess bezeichnet werden: die Nutzenden bewegen sich innerhalb einer bestimmten Anzahl von Wissenseinheiten, die sie wiederholt besuchen können und deren Verweildauer sie bestimmen (vgl. Kap. 6.6, *Sequenzen unterschiedlicher Länge*).

Im Zentrum der Ereignisdatenanalyse steht dabei die Kausal-Relation[67] von Ausgangszustand und Zielzustand: Das zentrale Anliegen der Kausal-Analyse kommt im Untertitel „New Approaches to Causal Analysis“ des Buches „Techniques of Event History Modelling“ von Blossfeld / Rohwer (2002) zum Ausdruck. Das Adjektiv „new“ betont dabei den Gegensatz zu „traditionellen“ Ansätzen der Analyse von Kausalbeziehungen anhand von Strukturgleichungsmodellen, in denen der Faktor Zeit keine Berücksichtigung findet

67 Zur Verwendung des Konzeptes „Kausalität“ im Rahmen der Ereignisanalyse, vgl. Blossfeld / Rohwer (2002: 21ff.).

(vgl. Blossfeld / Rohwer 2002: 22). Bei der Ereignisanalyse wird ein spezifisches Ereignis als *unabhängige Variable* definiert, um auf der Grundlage stochastischer Übergangswahrscheinlichkeiten zu analysieren, welchen Effekt dieses spezifische Ereignis auf weitere *abhängige Variablen* ausübt. Auf diese Weise untersucht die Ereignisdatenanalyse bedingende Faktoren spezifischer Ereignisse, beispielsweise in Form von verlaufsprägenden Ereignissen oder Wendepunkten („turning points"). Die generelle Forschungsperspektive der Ereignisdatenanalyse fasst Windzio (2001: 191) daher wie folgt zusammen: „Ziel der ereignisanalytischen Modelle ist die Schätzung der Einflüsse von Kovariaten auf die als abhängige Variable definierte Übergangsrate." Die Analyse von Ereignissen in Hinblick auf Kausalität stellt dabei eine besondere methodologische Herausforderung dar, wobei strikt zwischen *Aussagen über Ursachen (Kausalität)* und *Aussagen über Zusammenhänge (Korrelationen)* unterschieden wird. Kausalbeziehungen enthalten Aussagen darüber, wie Ereignisse entstehen und durch weitere Ereignisse beeinflusst werden und beziehen sich damit sowohl auf die Erklärung des Auftretens von Ereignissen als auch auf deren Vorhersage. Eine zentrale Bedeutung kommt bei der Kausalanalyse der Berücksichtigung der zeitlichen Reihenfolge des Auftretens spezifischer Ereignisse zu (die in dieser Form bei Aussagen über Korrelationen keine Bedeutung hat).

Ein entscheidender Punkt der Analyse von Kausalbeziehungen im Rahmen der Ereignisanalyse ist, dass diese ein *theoretisches Modell* von Kausalbeziehungen erfordert in Form von Hypothesen, auf welche Art und Weise eine spezifische Ursache zu einem spezifischen Effekt führt.[68] Es werden die Wahrscheinlichkeiten von Zustandswechseln in einem Verlauf analysiert (Übergangswahrscheinlichkeiten, Übergangsraten), um theoretisch entwickelte Modelle empirisch zu überprüfen (vgl. Rohwer / Trappe 1997). Implizit wird bei diesem Vorgehen von typischen (auf theoretischer Grundlage entwickelten) Mustern bzw. Verläufen ausgegangen.

Forschungsmethodologisch ergibt sich aus dieser Vorgehensweise jedoch die Schwierigkeit, dass in der Empirie vorhandene Muster, die im Vorfeld der Analyse nicht theoretisch entworfen wurden, auf der Grundlage der Methode der Ereignisdatenanalyse nicht entdeckt werden können: Die Ereignisdatenanalyse ist generell ein *hypothesengeleitetes Verfahren,* bei dem die zu analysierenden Ereignisse in eine kausale Beziehung gesetzt werden. Die Ereignisanalyse bildet diese kausalen Beziehungen jedoch nicht direkt ab. „The crucial point in regard to causal statements is, however, that they need a *theoretical argument* specifying the *particular mechanism of how a cause produces an effect*, or, more generally, *in which way interdependent forces affect each other in a given setting over time*. Therefore, the important task of event history modelling is not to demonstrate causal processes directly, but to establish relevant empirical evidence that can serve as a link in a chain of reasoning about causal mechanisms" (Blossfeld / Rohwer 2002: 24; Hervorhebung im Original). Darüber hinaus wird in der Ereignisanalyse eine grundsätzliche *Nichtdeterminiertheit des Handelns von Akteuren* postuliert, die auch theoretisch als grundsätzliche Unbestimmtheit in der Kausalanalyse

68 In Hinblick auf die empirische Überprüfung theoretischer Modelle kann die Ereignisanalyse als *konfirmatorische Analysestrategie* bezeichnet werden, vgl. Kap. 6.7: 57.

berücksichtig wird (vgl. Blossfeld / Rohwer 2002: 28). Diese Unbestimmtheit kommt gerade im Begriff der „Übergangs*wahrscheinlichkeit"* zum Ausdruck, die auf stochastischer Grundlage errechnet wird und sich darin von deterministischen Ansätzen und Modellen unterscheidet: „Übergangs*wahrscheinlichkeit"* wird definiert als die *Neigung* („propensity", Blossfeld / Rohwer 2002: 33) zur Veränderung des Zustandes von einem Ausgangszustand *J* zu einem Zielzustand *K* zum Zeitpunkt *t*. Die Ereignisanalyse basiert damit auf stochastischen Prozessen und fokussiert das Auftreten potentieller, zukünftiger Ereignisse und nicht deren konkrete Realisation.

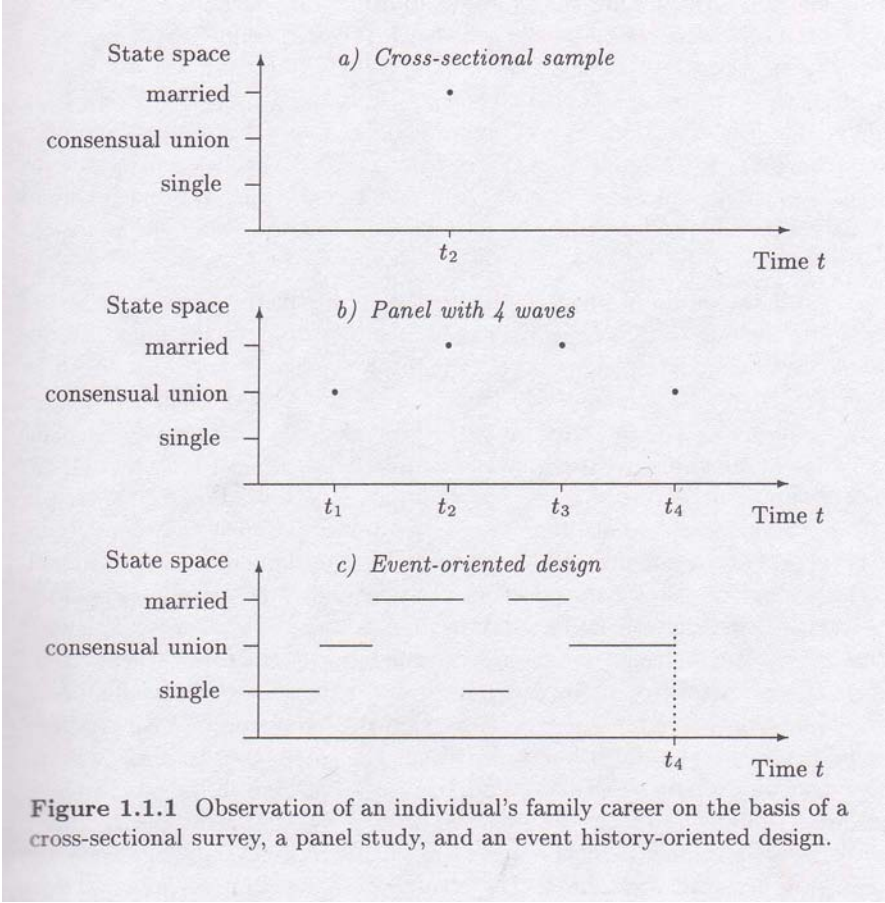

Figure 1.1.1 Observation of an individual's family career on the basis of a cross-sectional survey, a panel study, and an event history-oriented design.

Abbildung 14: Unterschiedliche Designs der Datenerhebung (Blossfeld / Rohwer 2002: 5)

Wie oben erläutert bilden den Gegenstand der Ereignisanalyse *Prozesse*, die als *diskrete Zustände* in einem *kontinuierlichen Zeitverlauf* abgebildet werden („continuous-time, discrete state substantive process" Blossfeld / Rohwer 2002: 4). Diese Fokussierung macht ein spezielles *Design der Datenerhebung* erforderlich, ebenso wie eine spezielle *Analyse und Interpretation* der erhobenen Daten. Abbildung 14 veranschaulicht drei prototypische Designs der Datenerhebung: Querschnittsdesigns („cross-sectional sample"), Längsschnittdesigns („panel") und ereignisorientierte Designs („event-oriented designs").[69] Nach Blossfeld und

69 Für eine ausführliche methodologische Diskussion der Vor- und Nachteile der unterschiedlichen Untersuchungsdesigns, vgl. Blossfeld / Rohwer (2002: 5ff.)

Rohwer (2002) bildet das ereignisorientierte Erhebungsdesigns die einzig angemessene Voraussetzung zur Analyse vieler empirischer Prozesse sowie deren Veränderungen: Grundlegende Kennzeichen sind die kontinuierliche Erhebung *qualitativer Variablen* sowie deren *zeitliche Verläufe*.

Das ereignisorientierte Design ermöglicht die Erhebung „most complete data possible on changes in qualitative variables that may occur at any point in time“ (Blossfeld / Rohwer 2002: 19). Die Daten der Ereignisanalyse werden in der Regel *retrospektiv* erhoben.[70] Im Gegensatz zu am Querschnitt- oder Längsschnittdesign orientierten Methoden zur Erhebung von *Zuständen* liegt der Schwerpunkt dabei auf der Erhebung von *Ereignissen* mit explizit zeitlichem Bezug. Dieser zeitliche Bezug kommt besonders deutlich in der Unterscheidung von „panel“ und „event-oriented design“ zum Ausdruck (vgl. Abbildung 14:*Unterschiedliche Designs der Datenerhebung (Blossfeld / Rohwer 2002: 5)*): Während beim „panel design“ als klassischer Längsschnittstudie Daten zu unterschiedlichen Zeitpunkten (t1 bis t4) erhoben werden, werden beim „event-oriented design“ der Zeitpunkt des Eintritts- und Austritts unterschiedlicher Zustände erhoben.

Eine methodenimmanente Schwierigkeit der Ereignisanalyse ist das „Zensierungsproblem“ (vgl. Abbildung 10: 56). Nicht bei allen Untersuchungseinheiten treten die im Fokus der Analyse stehenden Übergänge im Beobachtungszeitraum auf. Bei diesen „zensierten“ Zeitintervallen ist die Dauer bis zum Auftreten des Ereignisses unbekannt.[71] Bei der Ereignisdatenanalyse kann nach Elder (1985, vgl. Sackmann / Wingens 2001) grundsätzlich eine prospektive von einer retrospektiven Analyseperspektive unterschieden werden (vgl. Abb. 15): Den Referenzpunkt der *prospektiven* Analyseperspektive bildet ein gemeinsam geteilter Ausgangszustand, der auf unterschiedlichen Wegen verlassen wird. Den Referenzpunkt der *retrospektiven* Analyseperspektive bildet ein gemeinsam geteilter Endzustand, der auf unterschiedliche Art und Weisen erreicht wird.

In der soziologischen Lebenslaufforschung kommt durch die Verwendung der Ereignisdatenanalyse eine *dynamische Perspektive* auf Gesellschaft und auf soziale Phänomene zum Ausdruck, die grundsätzlich nicht als Eigenschaften von Individuen oder sozialen Gruppen interpretiert werden, sondern vielmehr als auf spezifische *Prozesse* verweisend. Eine besondere Bedeutung kommt dabei den komplexen und vielfältigen Prozessen des Eintretens, des Verweilens und des Verlassens von Zuständen zu. Diese spezifische dynamische Perspektive beschreiben Sackmann / Wingens (2001: 11) wie folgt: „Arbeitslosigkeit etwa ist in individueller, prozesshafter Sicht ein Zustand, in den eine Person gerät, der einige Zeit andauert und dann meist wieder verlassen wird. Für die politische und praktische Bearbeitung eines sozialen Problems wie z.B. Arbeitslosigkeit macht es nun einen ganz entscheidenden Unterschied, ob diese Bearbeitung als Behandlung von 'Problemgruppen' stattfindet oder ob diese Bearbeitung in Form einer Beeinflussung des Prozesses des indi-

70 Zur Diskussion der impliziten Begrenzungen der retrospektiven Form der Datenerhebung, vgl. Blossfeld / Rohwer (2002: 19f.), vgl auch Kap. 2: 5; *Methodologische Grundlagen der Navigationsanalyse*.

71 Zum Umgang mit diesem „Zensierungsproblem“ im Rahmen der Ereignisanalyse, vgl. Blossfeld / Rohwer (2002).

viduellen Eintritts in diesen, der Verweildauer in oder des individuellen Austritts aus diesem problematischen Zustand erfolgt."

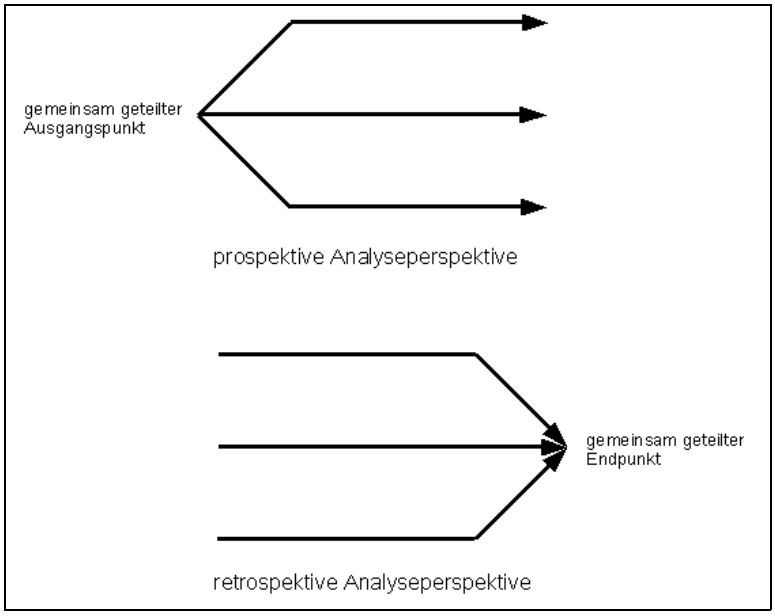

Abbildung 15: Prospektive und retrospektive Analyseperspektive (nach Elder 1985)

Als konzeptionelle Leitbegriffe der Lebenslaufperspektive nennen Sackmann / Wingens (2001: 12f.) „Übergang" und „Verlauf", die sie mit dem Begriff der „Sequenz" zu einer „grundbegriffliche Trias" erweitern: „Sequenz" ist als Konzept zwischen „Übergang" als einem singulären Statusübergang und „Verlauf" als Gesamtverlauf angesiedelt und wird definiert als Teil eines Gesamtverlaufs, „[...] der mindestens zwei Übergänge im Sinne von Zustandswechseln umfasst." (Sackmann / Wingens 2001: 32, vgl. auch Erzberger 2001, Erzberger / Prein 1997). Als Beispiele für grundlegende Typen von Sequenzen in der Lebenslaufforschung nennen Sackmann und Wingens (2001: 33-36) „Zwischenstatus", „Wechselstatus", „Brückenstatus", „Folgestatus" und „Statusabbruch". Ausdrücklich ist an dieser Stelle darauf hin zu weisen, dass mit der Bestimmung von „Sequenz" als aus mindestens zwei aufeinander folgenden Übergängen in einer Prozesszeit eine „Sequenz" aus der Abfolge von mindestens drei Zuständen definiert ist. In dieser Logik stellt eine Abfolge von zwei aufeinander folgenden Zuständen (d.h. einem Zustandswechsel) einen „Übergang" dar (vgl. Tabelle 8: 92).

Diese Bestimmung des Konzeptes „Sequenz" bildet einen Schnittpunkt von Ereignisanalyse und Sequenzanalyse, bei dem die zentrale Bedeutung der Abfolge aufeinander folgender Zustände im Sinne eines *vorher – nachher* auf der Basis einer grundlegenden zeitlichen Achse zum Ausdruck kommt. Der wesentliche Unterschied der Sequenz- und der Ereignisdatenanalyse besteht neben diesen terminologisch-konzeptionellen Differenzen in methodologischen Differenzen sowie in Differenzen des grundlegenden Analysegegenstandes (vgl. Tabelle 9: 94). Die Unterschiede im Hinblick auf den Gegenstand der jeweiligen Analyse hebt

auch Abbott (1990: 390) hervor: „The chief difference lies in their taking histories as wholes rather than as stepwise processes generated by iterated models."

- Den *Gegenstand der Sequenzdatenanalyse* bildet die vollständige, zusammenhängende Abfolge von Sequenzen (z.B. als Teil eines Gesamtverlaufs); die Analyse beruht auf Levenshtein-Distanzen als Ergebnis des paarweisen Vergleichs aller Sequenzen.
- Den *Gegenstand der Ereignisdatenanalyse* bildet dagegen die in einzelne, dyadische Übergänge zerlegte Sequenz: Auf stochastischer Grundlage werden jeweils die Übergange zwischen *einzelnen* Zuständen, sowie deren Zusammenhang analysiert. Die vollständige Sequenz als Gesamtverlauf kommt lediglich als Zusammenhang jeweils einzelner Übergänge in den Blick (Sackmann / Wingens 2001: 40).

Ereignis (event)	Wechsel zwischen zwei Zuständen zu einem bestimmten Zeitpunkt t.
Sequenz (sequence)	Abfolge von mindestens zwei aufeinander folgenden Übergängen in einer Prozesszeit.
Trajekt (trajectory)	Dichte, sequentielle, konditionelle und nicht intentionale Verkettung von Ereignissen, die mit dem Erleben eines Verlustes an Handlungskompetenz einhergeht.
Verlauf (trajectory)	Gesamtheit aller Übergänge und Verweildauern in Zuständen von einem Akteur. Je nach Fragestellung können bestimmte thematische Verläufe (z.B. Wohngeschichte) im Mittelpunkt des Interesses stehen (wobei andere thematische Verläufe wie z.B. Erwerbs- und Familiengeschichte als Randbedingung berücksichtigt werden).
Übergang (transition)	Wechsel zwischen zwei Zuständen in einem Prozess, der mehr oder weniger Zeit in Anspruch nimmt.
Übergangsstruktur (linkage structure)	Ensemble mehr oder weniger institutionalisierter gesellschaftlicher Verknüpfungen zwischen zwei Zuständen.
Wendepunkt (turning point)	Wechsel der Richtung eines Verlaufs innerhalb eines bestimmten Zeitraums. Als Richtungswechsel wird dabei ein Abbruch der Fortsetzung eines sozialstrukturell als wahrscheinlich zu erachtenden Verlaufspfades angesehen.
Zustand (state)	Eine für den jeweiligen Forschungsgegenstand zu definierende, in der Regel veränderbare Lebenslaufposition oder Eigenschaft innerhalb eines finiten Zustandsraums.
Zustandsraum (state space)	Eine definierte Menge sich gegenseitig ausschließender Lebenslaufpositionen oder Eigenschaften, die ein Akteur einnehmen kann.

Tabelle 8: Glossar zentraler Begriffe der Lebenslaufperspektive (Sackmann / Wingens 2001: 42)

8.1 Exkurs: Markov-Ketten

Im Folgenden wird anhand des *stochastischen Markov Prozesses* (Markov 1912; vgl. Windzio 2001) das grundlegende methodologische Modell der Ereignisdatenanalyse näher erläutert, um das ereignisanalytische Modell vom Modell der Sequenzdatenanalyse mittels Optimal-Matching abzugrenzen.

Allgemein besteht der zentrale Forschungsgegenstand von Markov-Prozessen in der Analyse der Abfolge von Zuständen durch die Berechnung von Übergangswahrscheinlichkeiten: Mit welcher Wahrscheinlichkeit folgt auf den Zustand A der Zustand B? Auf den Zustand B der Zustand C? Grundlegende Elemente von Markov-Ketten ist der *Zustandsraum* (als nichtleerer, endlicher Menge) und eine *stochastische Matrix*, die die jeweilige Wahrscheinlichkeit enthält, von einem spezifischen Zustand in einem Schritt in einem Folgezustand überzugehen.

Als *Markov-Prozesse erster Ordnung* (oder auch Markov-*Ketten* erster Ordnung) werden genau solche Prozesse bezeichnet, bei denen das Auftreten *folgender* Zustände lediglich vom *momentanen* Zustand abhängt – und nicht von *vorangehenden* Zuständen beeinflusst wird. Analyseeinheit ist der isolierte dyadische Übergang von Zuständen. Damit wird eine *Gedächtnislosigkeit* des Prozesses postuliert: die Eintrittswahrscheinlichkeit eines Zustandswechsels (Übergangswahrscheinlichkeit) eines Markov-Prozesses wird *nicht* von dessen *Vorgeschichte* beeinflusst und kann demzufolge *unabhängig* von den vorangehenden Zuständen prognostiziert werden.[72] Mit anderen Worten: zusätzliche Informationen über die Vergangenheit des Prozesses in Form vorangehender Zustände verbessern dabei *nicht* die Prognose der folgenden Zustände. Darüber hinaus ist die Eintrittswahrscheinlichkeit unabhängig von der Verweildauer in den vorangehenden Zuständen (vgl. Windzio 2001).

Diese *Markov-Prozesse erster Ordnung* werden durch das Konzept von *Markov-Prozessen zweiter Ordnung* erweitert, die auch als Semi-Markov-Prozesse bezeichnet werden. Die Erweiterung besteht darin, dass bei Markov-Ketten zweiter Ordnung nicht ausschließlich der momentane Zustand zur Prognose des folgenden verwendet wird, sondern eine *begrenzte Anzahl* vorangehender Zustände. Das Postulat der *Gedächtnislosigkeit* des Markov-Prozesses wird damit erweitert zur Berücksichtigung von *Vergangenheit*. Mit der Berücksichtigung einer bestimmten Anzahl vorangehender Zustände wird Prozessen Rechnung getragen, die nicht als *gedächtnislos* im Hinblick auf den Prozess*verlauf* oder die Prozess*zeit* betrachtet werden können. Gemeinsam sind beiden Konzepten jedoch die stochastische Grundlage und die Analyse dyadischer Übergänge, auch wenn diese bei Markov-Ketten zweiter Ordnung um eine begrenzte Anzahl vorangehender Zustände erweitert werden.

72 Vgl. den historischen Entwicklungskontext der Markov-Prozesse als statistisches Werkzeug zur Berechnung von Buchstabensequenzen (vgl. Markov 1912).

Zusammenfassend kann festgehalten werden, dass das Potenzial der Ereignisdatenanalyse in der Analyse von Determinanten von Übergängen, in der empirischen Überprüfung von Hypothesen und Kausalmodellen mit der Berücksichtigung parallel ablaufender Prozesse in Form von zeitveränderlicher Kovariaten besteht.[73]

Kritisiert werden an der Ereignisdatenanalyse in erster Linie methodenimmanente Grenzen wie die Fokussierung auf isolierte Ereignissen des Lebenslaufs als Analyseeinheit, wodurch der *Gesamtverlauf* ausgeblendet bleibt (Windzio 2001: 169; Abbott 1995a: 105; Aisenbrey 2001: 116; Han / Moen 1999: 197; Halpin / Chan 1998). Diese methodologische Ausrichtung stößt in genau den Anwendungsfällen der Ereignisanalyse an Grenzen, „[...] wenn nicht mehr nur der isolierte *Übergang* von einem Ausgangs- in einen Zielzustand, sondern der *gesamte Lebenslauf* betrachtet wird. Geht man von einem Verlaufspfad aus, bei dem auch bereits länger zurück liegende Ereignisse eine gegenwärtige Situation beeinflussen, ist eine auf Statusübergänge beschränkte Perspektive unter Umständen nicht mehr hinreichend“ (Windzio 2001: 163; Hervorhebung im Original).

Für die Analyse von Verläufen schlägt Windzio (2001) das Optimal-Matching Verfahren vor, da dieses gerade mit der Analyse vollständiger Sequenzen ein *Prozessgedächtnis* darstellt, im Gegensatz zur *Gedächtnislosigkeit* von Markov-Prozessen. Generell spricht Windzio in Abgrenzung zum stochastischen Markov-Prozess von Non-Markov-Prozessen, die das genaue Gegenteil von Markov-Prozessen darstellen: diese Prozesse können gerade *nicht* als *gedächtnislos* interpretiert und analysiert werden, sondern sind abhängig von der Verweildauer in den betreffenden Zuständen und von ihrer *Vorgeschichte*.

Sequenzdatenanalyse	**Ereignisdatenanalyse**
explorativ-heuristisch	hypothesengeleitet
deskriptiv	kausal-analytisch
Matrix mit Levenshtein-Distanzen	Matrix mit Übergangswahrscheinlichkeiten (stochastisch)
Analyse vollständiger Sequenz	Analyse dyadischer Übergänge

Tabelle 9: Gegenüberstellung Sequenzdatenanalyse - Ereignisdatenanalyse

Allgemein ist festzuhalten, dass es sich bei der Ereignisdatenanalyse und der Sequenzdatenanalyse um unterschiedliche Methoden handelt, mit spezifischen Zielen und spezifischen methodenimmanenten Potenzia-

73 Allerdings weisen Sackmann / Wingens (2001) darauf hin, dass das grundsätzliche, innovative Potenzial der Ereignisdatenanalyse – z.B. bei der Verwendung zeitveränderlicher Kovariaten - bisher erst wenig in der Forschungspraxis genutzt wird.

len und Grenzen. Windzio (2001: 164; Hervorhebung im Original) skizziert die spezifischen Ziele der Verwendung folgendermaßen: „Im Gegensatz zur Ereignisdatenanalyse, die eine adäquate Modellierung der *Determinanten von Übergängen* ermöglicht, liefern die deskriptiv-, explorativen Methoden der Sequenzdatenanalyse eine Voraussetzung für die Erstellung von *Typologien vollständiger Lebensläufe* oder Lebenslaufabschnitte.“ Das Potenzial der Kombination der Ereignis- und der Sequenzdatenanalyse betonen auch Sackmann / Wingens (2001: 41) und verweisen auf die Kombination des explorativ-heuristischen Vorgehens der Optimal-Matching Analyse mit dem hypothesengeleiteten Vorgehen der Ereignisdatenanalyse.

9 Navigationsanalyse als Sequenzdatenanalyse

In diesem Kapitel werden grundlegende methodologische Entscheidungen des Vorgehens der Navigationsanalyse als Sequenzdatenanalyse mittels Optimal-Matching begründet.
Die Grundlage dieser Begründung bildet die Analyse eines Teildatensatzes der vorliegenden empirischen Datenbasis (vgl. Kap. 10.1, *Datenerhebung und Datenbasis*). Es handelt sich dabei um die Sequenzen der Mikronavigation innerhalb der Lerneinheit „Maße der zentralen Tendenz“ (513), also um 475 Sequenzen mit 1542 Elementen (dies entspricht etwa einem Drittel des Umfangs der gesamten Datenbasis, vgl. Tab. 10: 126; *Anzahl der analysierten Wissenseinheiten und Sequenzen im Überblick*). Der Verwendung eines Teildatensatzes der Datenbasis liegt die Annahme zugrunde, dass die Effekte unterschiedlicher Entscheidungen in der Durchführung der Sequenzdatenanalyse mittels Optimal-Matching in einem formal und inhaltlich homogenen Bereich deutlicher erkennbar, inhaltlich interpretierbar und damit überprüfbar sind.
Zunächst wird die bei der Optimal-Matching Analyse verwendete *Definition der Substitutionskosten* empirisch begründet (Kap. 9.1). Dazu werden die drei in TDA implementierten Alternativen der Definition der Substitutionskosten (default-Definition der Substitutionskosten, datenbasierten Definition und Substitutionskosten als absolute Differenz) sowohl *formal* als auch *inhaltlich* analysiert und miteinander verglichen.
Daran anschließend wird das im Rahmen der Navigationsanalyse verwendete Clusterverfahren empirisch begründet (Kap. 9.2). Einführend werden hierarchische und partitionierende Verfahren der Clusteranalyse skizziert, sowie die grafische Darstellung des Agglomerationsprozesses durch Dendogramme. Die Clusterlösungen der in TDA implementierten Clusterverfahren werden sowohl *formal* beschrieben als auch *inhaltlich* miteinander verglichen.
Abschließend wird in diesem Kapitel die Entscheidung über die bei der Navigationsanalyse verwendete Clusteranzahl bzw. Fusionsebene begründet (Kap. 9.4).

9.1 Definition der Substitutionskosten im Rahmen der Optimal-Matching Analyse

Im Rahmen der Darstellung der Optimal-Matching Analyse wurde bereits auf die Gewichtung von Operationen durch *Kosten* als einem der zentralen Faktoren hingewiesen. Die Definition der Substitutionskosten hat direkten Einfluss auf die Berechnung der Distanz zwischen Sequenzen und somit auf die Distanzmatrix. Diese Levensthein-Distanzmatrix als Ergebnis der Optimal-Matching Analyse bildet dann die Grundlage

weiterer Analysen (*strukturen-prüfend* bzw. *strukturen-entdeckend*, vgl. Backhaus et al. 2000, s. auch Kap. 6.7: 57; *Potential der Optimal-Matching Analyse*).

Da für den Bereich der Analyse von Navigationsprozessen bislang keine Erfahrungen über die Effekte unterschiedlicher Definitionen von Substitutionskosten vorliegen, wird in diesem Kapitel die Definition der Substitutionskosten detailliert analysiert und begründet. Welche Auswirkung haben unterschiedlichen Definitionen der Substitutionskosten auf das Ergebnis der Optimal-Matching Analyse? Welche Definition von Substitutionskosten ist den vorliegenden Sequenzdaten angemessen?

```
seqm (
    m=...,      selection of method, def. 1
    sn=...,     selection of sequence data structure(s), def. 1
    icost=...,  indel cost specification
    scost=...,  substitution cost specification
```

Abbildung 16: Definition der Indel (icost) und der Substitutionskosten (icost) innerhalb des seqm-Befehls (Rohwer / Pötter 2005: 480)

Analysiert werden im Folgenden drei grundlegende und in TDA implementierte Varianten der Definition von Substitutionskosten.[74]

- *Grundeinstellung („default") der Substitutionskosten.* Die Werte der *default*-Kosten betragen für Indel „1" und für Substitution „2";
- *datenbasierte Definition der Substitutionskosten* (scost=2);
- *Substitutionskosten als absolute Differenz* der Sequenzen (scost=1).

Grundsätzlich ist bei der Definition von Substitutionskosten in TDA zwischen den Parametern des *seqpm*-Befehls (scost=1: Substitutionskosten als absolute Differenz; scost=2: datenbasierte Berechnung der Substitutionskosten) und dem *konkretem Wert* der Substitutionskosten zu unterscheiden.

Auf Grundlage des dargestellten Teildatensatzes werden je eine Optimal-Matching Analyse mit einer der Varianten der Substitutionskosten-Definitionen durchgeführt. Die daraus resultierenden Distanzmatrizen werden anhand der Ergebnisse einer Clusteranalyse miteinander verglichen. Die Kosten für die Operationen *Einfügen* („insertion") und *Löschen* („deletion") werden bei der folgenden Analyse auf den Wert „1" gesetzt und konstant gehalten.[75]

Das Vorgehen zur Beantwortung der oben gestellten Frage nach dem Einfluss der Definition der Substitutionskosten auf das Ergebnis der Optimal-Matching Analyse stellt sich folgendermaßen dar. Aufbauend auf

74 Die Option der Verwendung einer differenzierten, theoretisch begründeten Substitutionskostenmatrix wird aus inhaltlich-theoretischen Gründen nicht weiter verfolgt, vgl. Kap. 6.4: 50.

75 In Kapitel 6.5: 53 wurde bereits darauf hingewiesen, dass vor allem das Verhältnis der Indelkosten zu den Substitutionskosten für die Durchführung der Optimal-Matching Analyse von zentraler Bedeutung ist.

dem Ergebnis der Optimal-Matching Analyse unter jeweiliger Verwendung je einer der drei Varianten der Definition der Substitutionskosten wird eine Clusteranalyse nach Ward durchgeführt:

1. Dabei wird jeweils die *hierarchische Struktur des Agglomerationsprozesses* der letzten 28 Fusionsschritte anhand der entsprechenden Dendogramme verdeutlicht;
2. für die Clusterlösung mit 28 Clustern werden die *Häufigkeitsverteilung* als Zuordnung von Fällen zu Clustern dargestellt;
3. anschließend werden die *Korrelationen* (Cramer-V) der spezifischen Clusterlösungen berechnet: Wie ähnlich sind die Clusterlösungen, die auf unterschiedlichen Substitutionskosten beruhen?
4. Anhand von *Kreuztabellen* wird die Zuordnung von Fällen in Cluster analysiert: In welchen Zuordnungen besteht Übereinstimmung? In welchen Zuordnungen bestehen Differenzen? Diese Zuordnungen werden inhaltlicher Ebene interpretiert.
5. Als abschließendes Fazit folgt die *Begründung* der Wahl der Substitutionskosten-Definition im Rahmen der Navigationsanalyse.

9.1.1 Struktur der Agglomerationsprozesse

Die folgenden Dendogramme veranschaulichen die Struktur des Agglomerationsprozesses der hierarchischen Clusteranalyse (Ward) der letzten 28 Fusionsschritte und ermöglichen einen ersten Überblick über den Einfluss unterschiedlicher Substitutionskosten-Definitionen.

- Abbildung 17 zeigt den Agglomerationsprozess auf Grundlage der *default-Definition* der Substitutionskosten;
- Abbildung 18 zeigt den Agglomerationsprozess auf Grundlage der *datenbasierten Definition* der Substitutionskosten und
- Abbildung 19 zeigt den Agglomerationsprozess auf Grundlage der Errechnung der Substitutionskosten als *absoluter Differenz.*

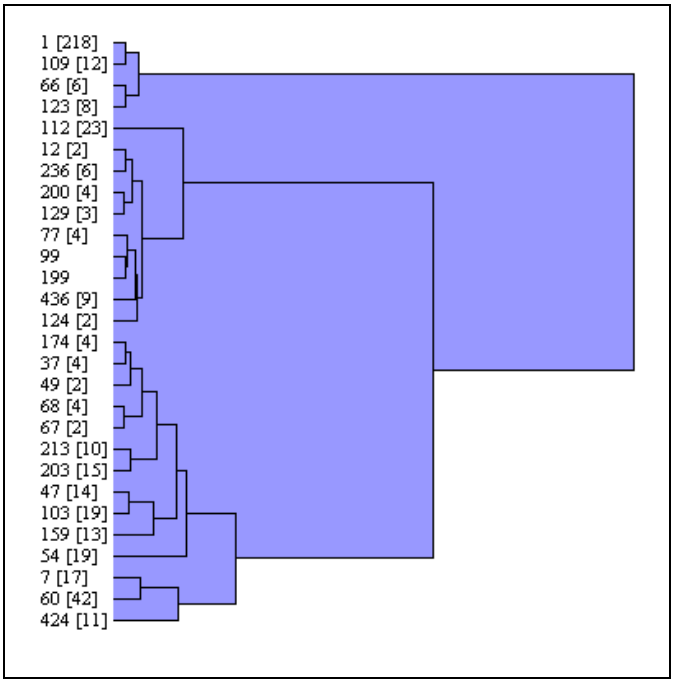

Abbildung 17: Ergebnis der Optimal-Matching Analyse (TDA): icost (Wert „1“) und scost (Wert „2“), entspricht den default-Einstellungen.

Bereits ein erster allgemeiner Vergleich der Dendogramme macht deutlich, dass sich die Struktur des Agglomerationsprozesses auf Grundlage der Substitutionskosten als absoluter Differenz und der *default*-Substitutionskosten ähneln. Die Struktur des Agglomerationsprozesses auf Grundlage der *default*-Substitutionskosten und der datenbasiert ermittelten Substitutionskosten zeigen dagegen weniger Ähnlichkeit.

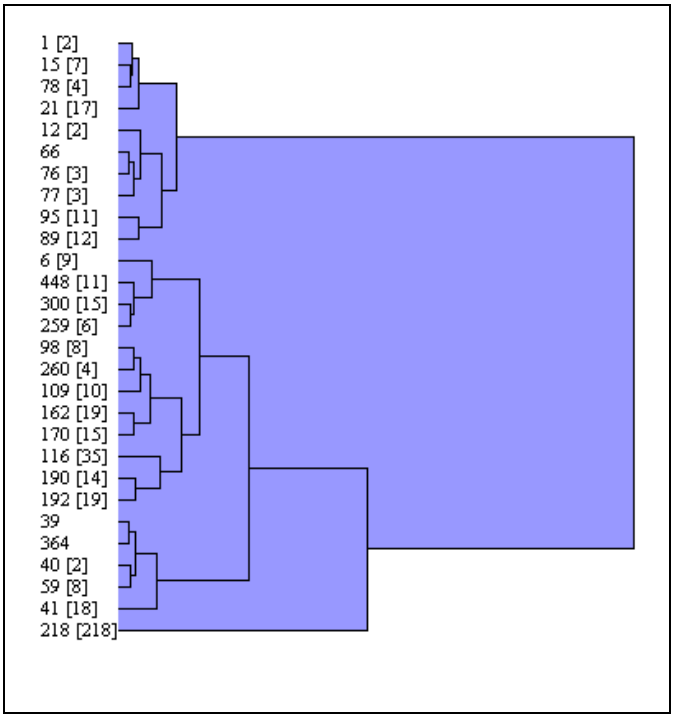

Abbildung 18: Ergebnis der Optimal Matching Analyse (TDA): icost (Wert „1“) und scost (Parameter „2“), d.h. Substitutionskosten werden datenbasiert ermittelt.

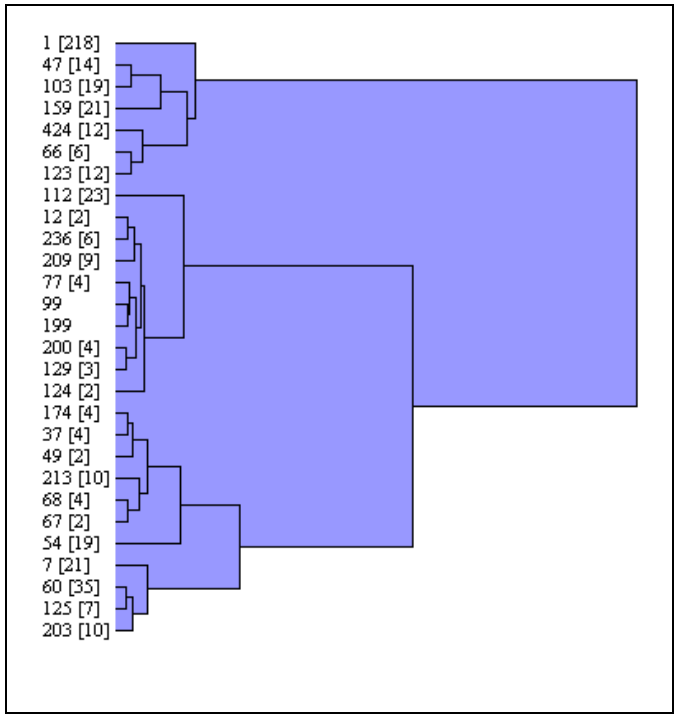

Abbildung 19: Ergebnis der Optimal-Matching Analyse (TDA): icost (Wert „1“) und scost (Parameter „1“), d.h. die Substitutionskosten werden errechnet als die absolute Differenz der Sequenzen.

In den folgenden Abschnitten werden die dargestellten Agglomerationsprozesse in *formaler* sowie *inhaltlicher* Hinsicht detailliert analysiert. Dazu werden im nächsten Abschnitt die *Häufigkeitsverteilungen* der Clusterlösungen dargestellt, d.h. die Anzahl der Fälle, die in die betreffenden Cluster fusioniert werden.

9.1.2 Häufigkeitsverteilungen der Clusterlösungen

Die folgenden Tabellen zeigen die Häufigkeitsverteilungen als detaillierte Darstellung des oben beschriebenen Agglomerationsprozesses für die Clusterlösungen mit 28 Clustern. Analog zur Darstellung anhand von Dendogrammen werden anhand dieser tabellarischen Darstellung Tendenzen von Gemeinsamkeiten und Differenzen der spezifischen Clusterlösungen erkennbar:

- Tabelle 20 zeigt die Häufigkeitsverteilung der Ward-Clusterlösung auf Grundlage der *default-Definition* der Substitutionskosten*;*
- Tabelle 21 zeigt die Häufigkeitsverteilung der Ward-Clusterlösung auf Grundlage der *datenbasierten Definition* der Substitutionskosten;
- Tabelle 22 zeigt die Häufigkeitsverteilung der Ward-Clusterlösung auf Grundlage der Definition der Substitutionskosten als *absoluter Differenz.*

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	218	45,9	45,9	45,9
	2	12	2,5	2,5	48,4
	3	6	1,3	1,3	49,7
	4	8	1,7	1,7	51,4
	5	23	4,8	4,8	56,2
	6	2	,4	,4	56,6
	7	6	1,3	1,3	57,9
	8	4	,8	,8	58,7
	9	3	,6	,6	59,4
	10	4	,8	,8	60,2
	11	1	,2	,2	60,4
	12	1	,2	,2	60,6
	13	9	1,9	1,9	62,5
	14	2	,4	,4	62,9
	15	4	,8	,8	63,8
	16	4	,8	,8	64,6
	17	2	,4	,4	65,1
	18	4	,8	,8	65,9
	19	2	,4	,4	66,3
	20	10	2,1	2,1	68,4
	21	15	3,2	3,2	71,6
	22	14	2,9	2,9	74,5
	23	19	4,0	4,0	78,5
	24	13	2,7	2,7	81,3
	25	19	4,0	4,0	85,3
	26	17	3,6	3,6	88,8
	27	42	8,8	8,8	97,7
	28	11	2,3	2,3	100,0
	Gesamt	475	100,0	100,0	

Abbildung 20: Häufigkeitsverteilung der Ward-Clusterlösung auf Grundlage der default-Definition der Substitutionskosten.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	2	,4	,4	,4
	2	7	1,5	1,5	1,9
	3	4	,8	,8	2,7
	4	17	3,6	3,6	6,3
	5	2	,4	,4	6,7
	6	1	,2	,2	6,9
	7	3	,6	,6	7,6
	8	3	,6	,6	8,2
	9	11	2,3	2,3	10,5
	10	12	2,5	2,5	13,1
	11	9	1,9	1,9	14,9
	12	11	2,3	2,3	17,3
	13	15	3,2	3,2	20,4
	14	6	1,3	1,3	21,7
	15	8	1,7	1,7	23,4
	16	4	,8	,8	24,2
	17	10	2,1	2,1	26,3
	18	19	4,0	4,0	30,3
	19	15	3,2	3,2	33,5
	20	35	7,4	7,4	40,8
	21	14	2,9	2,9	43,8
	22	19	4,0	4,0	47,8
	23	1	,2	,2	48,0
	24	1	,2	,2	48,2
	25	2	,4	,4	48,6
	26	8	1,7	1,7	50,3
	27	18	3,8	3,8	54,1
	28	218	45,9	45,9	100,0
	Gesamt	475	100,0	100,0	

Abbildung 21: Häufigkeitsverteilung der Ward-Clusterlösung auf Grundlage der datenbasierten Definition der Substitutionskosten

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	218	45,9	45,9	45,9
	2	14	2,9	2,9	48,8
	3	19	4,0	4,0	52,8
	4	21	4,4	4,4	57,3
	5	12	2,5	2,5	59,8
	6	6	1,3	1,3	61,1
	7	12	2,5	2,5	63,6
	8	23	4,8	4,8	68,4
	9	2	,4	,4	68,8
	10	6	1,3	1,3	70,1
	11	9	1,9	1,9	72,0
	12	4	,8	,8	72,8
	13	1	,2	,2	73,1
	14	1	,2	,2	73,3
	15	4	,8	,8	74,1
	16	3	,6	,6	74,7
	17	2	,4	,4	75,2
	18	4	,8	,8	76,0
	19	4	,8	,8	76,8
	20	2	,4	,4	77,3
	21	10	2,1	2,1	79,4
	22	4	,8	,8	80,2
	23	2	,4	,4	80,6
	24	19	4,0	4,0	84,6
	25	21	4,4	4,4	89,1
	26	35	7,4	7,4	96,4
	27	7	1,5	1,5	97,9
	28	10	2,1	2,1	100,0
	Gesamt	475	100,0	100,0	

Abbildung 22: Häufigkeitsverteilung der Ward-Clusterlösung auf Grundlage der Substitutionskosten als absoluter Differenz.

Auffällig ist bei den dargestellten Häufigkeitsverteilungen vor allem, dass in jeder der drei Clusterlösungen ein zahlenmäßig großes Cluster gebildet wird, das übereinstimmend aus 218 Fällen besteht. Eine detaillierte inhaltliche Analyse in Form von Kreuztabellen (vgl. Kap. 9.1.4) zeigt jedoch, dass diese Cluster zwar in der Anzahl der Fälle übereinstimmen, es sich aber um unterschiedliche Fälle handelt, die dem Cluster zugeordnet werden. Die detaillierte formale Analyse der unterschiedlichen Clusterlösungen wird im folgenden Abschnitt dargestellt.

9.1.3 Formale Analyse der unterschiedlichen Clusterlösungen

Zur detaillierten Analyse der oben dargestellten Häufigkeitsverteilungen werden die berechneten Clusterlösungen miteinander verglichen: Wie ähnlich sind die Clusterlösungen, die auf unterschiedlichen Substitutionskosten beruhen?

Da es sich bei der Zuordnung der Clusterzugehörigkeit zur je spezifischen Sequenz um nominalskalierte Daten ohne implizite Rangfolge handelt, werden im Folgenden die Korrelationen auf Grundlage von Cramer-V als Chi-Quadratbasiertes Zusammenhangsmaß errechnet.

- Abbildung 23 zeigt die Korrelation (Cramer-V) der Clusterlösungen der *default*-Definition und der *datenbasierten* Definition der Substitutionskosten;
- Abbildung 24 zeigt die Korrelation (Cramer-V) der Clusterlösungen für die *default*-Definition und der Definition der Substitutionskosten als *absoluter Differenz;*
- Abbildung 25 zeigt die Korrelation (Cramer-V) der Clusterlösungen für die *datenbasierte* und der Definition der Substitutionskosten als *absoluter Differenz.*

		Wert
Nominal- bzgl. Nominal-maß	Phi	1,369
	Cramer-V	,264
Anzahl der gültigen Fälle		475

a Die Null-Hyphothese wird nicht angenommen.
b Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Abbildung 23: Korrelation der Clusterlösungen (datenbasierten und der default-Substitutionskosten)

		Wert
Nominal- bzgl. Nominal-maß	Phi	5,004
	Cramer-V	,963
Anzahl der gültigen Fälle		475

a Die Null-Hyphothese wird nicht angenommen.
b Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Abbildung 24: Korrelation der Clusterlösungen (default- und Substitutionskosten als absoluter Differenz)

Symmetrische Maße

		Wert
Nominal- bzgl. Nominal-maß	Phi	1,354
	Cramer-V	,261
Anzahl der gültigen Fälle		475

a Die Null-Hyphothese wird nicht angenommen.
b Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Abbildung 25: Korrelation der Clusterlösungen (datenbasierten und der Substitutionskosten als die absolute Differenz)

Die Interpretation der dargestellten Korrelationen führt zu folgendem Ergebnis:

- Zwischen der *default*-Definition und der *datenbasierten* Definition der Substitutionskosten besteht ein schwacher Zusammenhang von ,264 (Cramer-V).
- Zwischen der *default*-Definition der Substitutionskosten und der Definition der Substitutionskosten als die *absoluter Differenz* besteht ein sehr starker Zusammenhang von ,963 (Cramer-V).
- Zwischen der *datenbasierten* Definition der Substitutionskosten und der Definition der Substitutionskosten als die *absoluter Differenz* besteht ein schwacher Zusammenhang von ,261 (Cramer-V).
- Die Clusterlösung auf Grundlage der *default*-Definition der Substitutionskosten ist nahezu identisch mit der Errechnung der Substitutionskosten als *absoluter Differenz* (Cramer-V: ,963).

Der Zusammenhang der Clusterlösungen auf Grundlage der Substitutionskosten als *absoluter Differenz* und *default*-Substitutionskosten ist mit .963 (Cramer-V) als extrem hoch und als nahezu identisch zu interpretieren: Diese Clusterlösungen unterscheiden sich lediglich durch die Zuordnung weniger Fälle.[76] Zur Erklärung dieser sehr geringen Unterschiede kann das 'bottom-up' Vorgehen der hierarchisch-agglomerativen Clusteranalyse herangezogen werden, mit der grundlegenden Eigenschaft der *Nichtrevidierbarkeit* der Clusterzuordnung (vgl. Kap. 9.2.1: 108).

Da die Clusterlösung der *default*-Definition der Substitutionskosten und der Substitutionskosten als *absoluter Differenz* nahezu identisch sind, beziehen sich die folgenden Schritte auf die Analyse der *default*- und der *datenbasierten* Definitionen der Substitutionskosten. Der Zusammenhang von .264 (Cramer-V) macht deutlich, dass diese beiden Definitionen der Substitutionskosten zu unterschiedlichen Clusterlösungen führen, die im folgenden Abschnitt inhaltlich analysiert werden.

9.1.4 Inhaltliche Analyse der unterschiedlichen Clusterlösungen

Unter Hinweis auf die schwache Korrelation der Clusterlösung auf Grundlage der default- und der datenbasierten Definition von Substitutionskosten (Cramer-V: .264) kann die Wahl der Substitutionskosten-Definition im Rahmen der Navigationsanalyse für die vorliegenden Daten nicht *formal*, sondern nur *inhaltlich* begründet werden. Dazu wird im Folgenden die Zuordnung der Fälle zu spezifischen Clustern anhand einer Kreuztabelle analysiert. Abbildung 26 stellt die Zuordnung der Fälle auf Grundlage der default-Definition (1. Zeile) und der datenbasierten Definition der Substitutionskosten (1. Spalte) tabellarisch dar.

76 Vgl. Anhang Kapitel 17.16, *Kreuztabelle der Clusterlösung auf der Grundlage der default-Definition der Substitutionskosten und der Substitutionskosten als absoluter Differenz*.

Diese *formale* Beschreibung der Zuordnung von Fällen zu Clustern wird im Folgenden *inhaltlich* interpretiert. Aufgrund dieser Interpretation wird dann die Entscheidung für die Verwendung der Substitutionskostendefinition getroffen.

Den Ausgangspunkt dieser inhaltlichen Analyse der Clusterlösungen stellt das

- auf der *default-Definition* der Substitutionskosten berechnete *Cluster 1*
- sowie das auf *datenbasierter* Definition der Substitutionskosten berechnete *Cluster 28* dar.

Die Kreuztabelle macht deutlich, dass beide Clusterlösungen aus jeweils 218 Fällen bestehen und zeigt darüber hinaus, dass in dieser Zuordnung lediglich eine Übereinstimmung von 86 Fällen besteht. Diese 86 Fälle stellen also die *Schnittmenge* der Cluster 1 und Cluster 28 dar.

Anzahl

		cg_subcostdatenb_ward_28c																											Gesamt
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
cg_subcost default_ward_28c	1	2	7	3	12	1	0	3	1	6	8	2	5	6	3	4	0	5	11	7	18	1	9	1	1	2	3	11	86	218
	2	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	0	0	2	0	0	0	0	0	0	0	6	12
	3	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	6
	4	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	1	0	1	0	0	0	0	0	0	0	3	8
	5	0	0	0	2	0	0	0	0	0	1	1	0	0	0	0	0	2	0	3	1	3	0	0	0	0	1	0	9	23
	6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
	7	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	2	6
	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	2	4
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	3
	10	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4
	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
	13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	2	1	0	0	0	0	0	4	9
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	2
	15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	4
	16	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	4
	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2
	18	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	4
	19	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
	20	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	1	1	4	10
	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0	0	0	9	15
	22	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	2	0	0	0	3	0	0	0	0	1	4	14
	23	0	0	0	0	0	0	0	0	0	1	0	0	2	0	1	0	0	2	3	1	0	1	0	0	0	0	0	8	19
	24	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0	0	0	1	0	7	13
	25	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2	13	19
	26	0	0	0	0	0	0	0	0	0	0	5	2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	8	17
	27	0	0	0	0	0	0	0	0	0	0	0	1	1	2	0	2	0	1	1	1	0	0	0	0	0	2	0	31	42
	28	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	1	6	11
Gesamt		2	7	4	17	2	1	3	3	11	12	9	11	15	6	8	4	10	19	15	35	14	19	1	1	2	8	18	218	475

Abbildung 26: Kreuztabelle der Zuordnung von Fällen zu Clustern (default- und datenbasierte Definition der Substitutionskosten)

Welche Clusterlösung ist nun aus *inhaltlicher* Perspektive als den Daten angemessener zu beurteilen? Dazu wird die *Schnittmenge* der beiden Clusterlösungen analysiert: Es wird also inhaltlich interpretiert, ob die Zuordnung dieser 86 Fälle zum Cluster 28 (datenbasiert) oder zum Cluster 1 (default) inhaltlich angemessener ist.

Die konkreten Sequenzen der Schnittmenge bestehen aus 86 identischen Fällen, dem einmaligen Aufruf der Wissenseinheit Orientierung (2642) innerhalb der Lerneinheit „Maße der zentralen Tendenz“ (513):

- Das auf *default*-Substitutionskosten berechnete Clustern 1 besteht ohne diese Schnittmenge aus 132 identischen Sequenzen. Aus inhaltlicher Perspektive entspricht dies dem einmaligen Aufruf der Wissenseinheit Orientierung (2642) innerhalb der Lerneinheit. Das Cluster 1 stellt somit eine sehr homogene Fusionierung von Fällen dar.

[...]
2642
2642
2642
2642
2642
2642
2642
2642
2642
2642
2642
2642
2642
[...]

Abbildung 27: Auszug: Fälle des Clusters 1 (default-Substitutionskosten) ohne Schnittmenge

- Das auf *datenbasierten Substitutionskosten* berechnete Cluster 28 besteht ohne die beschriebene Schnittmenge aus 132 *unterschiedlichen* Sequenzen (vgl. Abb. 28 für einen Auszug dieser 132 Sequenzen). Im Vergleich zur Clusterlösung auf Grundlage der *default*-Substitutionskosten ist das Cluster 28 als weniger homogen zu bezeichnen.

[...]
2896 2900
3365 3597 2900
3597 2900
3597 2900
3597 2900
3597 2900
3597 2900 2896
3597 2900 2896
2642 3597
2642 3597
2642 3597
2642 3597 2896
2642 3597 2642
2642 3597 2642
2642 3597 2896 3597 2642
2642 3597 2900
3597 2642
3597 2642
[...]

Abbildung 28: Auszug: Fälle des Clusters 28 (datenbasierte Substitutionskosten) ohne Schnittmenge

Die inhaltliche Analyse der Clusterlösungen auf Grundlage der *default*-Substitutionskosten und der datenbasierten Definition der Substitutionskosten führt zu dem Ergebnis, dass die Verwendung der default-Substitutionskosten zu einer deutlich *homogeneren* Clusterlösung führt.

9.1.5 Begründung der Wahl der Substitutionskosten-Definition

Da die Entscheidung für die Verwendung der Definition der Substitutionskosten nicht allein aufgrund *formaler* Kriterien getroffen werden kann, ist die Analyse der Clusterlösungen auf inhaltlicher Ebene notwendig. Im vorangehenden Abschnitt wurde gezeigt, dass lediglich eine schwache Übereinstimmung der Clusterlösung auf Grundlage der *default*-Substitutionskosten und der *datenbasierten* Definition der Substitutionskosten festzustellen ist.

Darüber hinaus wurde gezeigt, dass für die vorliegenden Daten die Verwendung der default-Substitutionskosten zu einem deutlich homogeneren Ergebnis führt als die Verwendung der datenbasierten Substitutionskosten: Daher werden für die Optimal-Matching Analyse im Rahmen der Navigationsanalyse die *default*-Substitutionskosten verwendet.

Dieses Ergebnis steht in Einklang mit theoretischen Überlegungen zur Gewichtung von Operationen durch *Kosten* (vgl. Kap. 6.3: 48) mit der Schlussfolgerung, dass im Rahmen der Navigationsanalyse keine Gewichtung von Substitutionen auf theoretischer Grundlage vorgenommen werden kann, da es keine als groß oder klein definierten Sprünge gibt. Die unterschiedlichen Navigationsweisen stehen ohne Gewichtung gleichberechtigt nebeneinander: es gibt keine Übergänge, die als teuer oder billig definiert werden können.

Darüber hinaus ist die Verwendung der *default*-Substitutionskosten transparenter und somit für den konkreten Fall besser nachvollziehbar als die Verwendung der *datenbasierten* Substitutionskosten: Wie bereits dargestellt beruht die *datenbasierte* Definition der Substitutionskosten auf der Berechnung der Übergangswahrscheinlichkeiten im vorhandenen Datensatz. Dies steht in gewissem Gegensatz zu dem grundsätzlichen methodologischen Vorgehen der Sequenzdatenanalyse, das sich ja gerade von der Berechnung von Übergangswahrscheinlichkeiten abgrenzt.

9.2 Clusteranalyse im Rahmen der Navigationsanalyse

In diesem Kaptitel wird das statistische Verfahren der Clusteranalyse als einem zentralen methodischen Bestandteil der Navigationsanalyse dargestellt. Auf dieser Grundlage werden im folgenden Kapitel (Kap. 9.3) die in TDA implementierten Clusterverfahren auf formaler Ebene beschrieben und analysiert. Das Kapitel ist wie folgt aufgebaut: Einleitend wird das Verfahren der Clusteranalyse unter Berücksichtigung hierarchischer (Kap. 9.2.1) sowie partitionierender Verfahren (Kap. 9.2.2) skizziert, daran anschließend wird die grafische Darstellung des Agglomerationsprozesses anhand von Dendogrammen dargestellt (Kap. 9.2.3).

9.2.1 Hierarchische Verfahren der Clusteranalyse

In diesem Abschnitt wird die im Rahmen der Navigationsanalyse verwendete Methode der Clusteranalyse näher erläutert (vgl. Bortz 1999, Wishart 1999, Kaufman / Rousseeuw 2005). Den Ausgangspunkt des Verfahrens der Clusteranalyse bildet im Rahmen der Navigationsanalyse die auf Grundlage der Optimal-Matching Analyse erstellte Levensthein-Distanzmatrix. Ausgehend von den Levenshtein-Distanzen als Ergebnis des paarweisen Sequenzvergleichs werden die Sequenzen des Datensatzes gruppiert.

Generell stellt die Clusteranalyse ein *heuristisches, strukturen-entdeckendes Verfahren* dar, um Objekte *systematisch* zu *Klassifizieren*. Die Objekte werden dabei aufgrund *definierter Merkmale* in *Gruppen (Cluster)* eingeteilt. Voraussetzung und Ausgangspunkt einer Clusteranalyse ist die Definition eines *Maßes*, mit dem die Ähnlichkeit bzw. Unähnlichkeit (Distanz) zwischen Objekten numerisch ausgedrückt wird. Im Fall der Navigationsanalyse ist dieses Maß die Levenshtein-Distanz zwischen Sequenzen.

Generelles Ziel der Clusteranalyse ist es, die *beste* Struktur der Objekte in Form der *besten* Aufteilung in Gruppen zu entdecken: die Cluster sind nach innen (intern) möglichst *homogen* und gleichzeitig nach außen (extern) möglichst *heterogen*. Dieses generelle Ziel fasst Bortz (1999: 547) folgendermaßen zusammen: „Mit der Clusteranalyse werden die untersuchten Objekte so gruppiert, daß die Unterschiede zwischen den Objekten einer Gruppe bzw. eines 'Clusters' möglichst gering und die Unterschiede zwischen den Clustern möglichst groß sind."

Allgemein wird die Bezeichnung „Clusteranalyse" als Sammelbegriff für eine Vielzahl unterschiedlicher Vorgehensweisen verwendet, die konkret auf unterschiedlichen Clusteralgorithmen basieren. Allgemein können bei Clusteranalysen hierarchische und nicht-hierarchische Vorgehensweisen unterschieden werden. Hierarchische Clusteranalysen lassen sich weiter in agglomerative und divisive Clusterverfahren differenzieren.

Bei *hierarchisch-agglomerativen* Clusterverfahren bildet als Ausgangspunkt jedes Objekt ein eigenes Cluster (vgl. Wishart 1999: 24). In einem ersten Schritt werden die beiden Objekte gruppiert, die die größte Ähn-

lichkeit aufweisen (dieses Gruppieren wird auch als „Fusionieren“ bezeichnet). Diese beiden Fälle werden in einem Cluster zusammengefasst. Daran anschließend werden die Abstände zwischen dem neu gebildeten Cluster und allen anderen Fällen aktualisiert. Es werden erneut die Objekte fusioniert, die die größte Ähnlichkeit aufweisen.

Die Clusteranalyse stellt somit einen *iterativen Prozess* dar: es werden diejenigen beiden Fälle in ein Cluster fusioniert (1. Fusionsschritt), die in der gesamten Objektmenge die kleinste Distanz aufweisen, sich also am ähnlichsten sind. Die Anzahl der Cluster insgesamt verringert sich damit um 1. Im nächsten Schritt werden wieder die beiden Cluster mit der geringsten Distanz fusioniert. Im letzten Schritt werden alle Fälle in einem einzigen Cluster zusammengefasst. Agglomerative Clusterverfahren folgen somit einer 'bottom-up'-Strategie. Das Vorgehen *hierarchisch-divisiver* Clusterverfahren folgt einer 'top-down'-Strategie. Als Ausgangspunkt befinden sich alle Objekte in einem Cluster, das dann schrittweise in kleinere Cluster unterteilt wird. Innerhalb der Gruppe der *hierarchisch-agglomerativen* Verfahren der Clusteranalyse bestehen Unterschiede der Clusteralgorithmen in der konkreten Definition des Konzeptes von „Ähnlichkeit“ bzw. „Distanz“. Gerade diese Definition des „Abstandes“ zwischen Objekten / Clustern ist ausschlaggebend für die Fusionsschritte im Rahmen der Clusteranalyse.

Im Folgenden werden die in TDA implementierten Clusteralgorithmen kurz beschrieben (vgl. Wishart 1984, 1999; Bortz 1999, Kaufman / Rousseeuw 2005):[77]

- Minimum-Methode („single linkage“[78]): Der Abstand von Clustern wird bestimmt durch den minimalen Abstand zweier Elementen aus den jeweiligen Clustern. Fusioniert werden die beiden Cluster mit den am nächsten zueinander liegenden Elementen.
- Maximum-Methode („complete linkage“): Der Abstand von Clustern wird bestimmt durch die jeweils am weitesten entfernten Objekte der jeweiligen Cluster („furthest neighbour“), d.h. der Objekte mit maximalem Abstand. Fusioniert werden die Cluster, für die diese maximale Distanz am geringsten ist.
- *Average linkage* („group average“): Der Abstand von Clustern wird bestimmt durch den Mittelwert der Abstände der Objekte des jeweiligen Clusters. Fusioniert werden die Cluster mit dem kleinsten durchschnittlichen Mittelwert.

  Eine Erweiterung des *Average Linkage* Verfahrens stellt das *weighted average linkage*-Verfahren dar: Dabei werden die durchschnittlichen Distanzen *gewichtet*, und zwar aufgrund der Anzahl der Objekte des spezifischen Clusters.
- Medianverfahren („centroid method“): Fusioniert werden die Cluster, deren Schwerpunkt als quadrierter, euklidischer Centroidabstand minimal ist. Der Centroidabstand steht dabei für die durchschnittlichen Merkmalsausprägungen aller Objekte des entsprechenden Clusters (vgl. Bortz 1999:

77 Eine Kurzbeschreibung der in ClustanGraphics implementierten Clusteralgorithmen befindet sich in Anhang, Kap. 17.1 (Auszug der ClustanGraphics Hilfe-Datei).

78 In SPSS ist das „single linkage“-Verfahren unter der Bezeichnung „nearest neighbour“ implementiert.

555). Die Objekthäufigkeiten der zu fusionierenden Cluster werden bei diesem Verfahren nicht berücksichtigt.

Eine Erweiterung des Medianverfahrens stellt das *gewichtete Centroid* bzw. *gewichtete Median*-Verfahren dar, bei dem unterschiedliche Objekthäufigkeiten berücksichtigt werden.

- Ward-Verfahren („Minimum-Varianz-Methode“, „Increase in Sum of Squares“, „Fehlerquadratsummen-Methode“, „HGROUP-100 Methode“): In einem iterativen Prozess werden genau die beiden Cluster fusioniert, deren Fusion die geringste Erhöhung der Binnenvarianz als der gesamten Fehlerquadratsumme darstellt (zur detaillierten Darstellung des Ward-Algorithmus, vgl. Bortz 1999, Wishart 1984, Wishart 1999).

Eine grundlegende methodische Schwäche hierarchischer Clusteranalysen liegt in dem beschriebenen Vorgehen der Fusionierung von Objekten begründet: Wird ein Objekt im iterativen Prozess der Clusteranalyse einem Cluster zugeordnet, bleibt diese Zuordnung im gesamten weiteren Verlauf der Clusteranalyse bestehen und kann nicht revidiert werden (vgl. Bortz 1999: 554). Allgemein wird daher empfohlen, die aufgrund der hierarchischen Clusteranalyse identifizierten Cluster mit Hilfe eines nicht-hierarchischen Verfahrens zu überprüfen und zu optimieren (z.B. mit Hilfe des nicht-hierarchischen *k-Means* Verfahrens).

9.2.2 Partitionierende Clusterverfahren

Den Ausgangspunkt *nicht-hierarchischer, partitionierender* Clusterverfahren bildet eine Startgruppierung, d.h. eine vorab definierte Anzahl von Clustern. Durch schrittweises Verschieben einzelner Fälle zwischen den Clustern wird versucht, die Homogenität der Cluster zu erhöhen und somit die Qualität der Zuordnung zu optimieren. Partitionierende Clusterverfahren sind beendet, wenn durch ein weiteres Verschieben von Fällen zwischen den Clusterlösungen keine Verbesserung der Homogenität erreicht werden kann. Die nicht-hierarchischen, partitionierenden Verfahren beruhen also auf einer iterativ-partiellen Vorgehensweise. Wie auch bei den hierarchischen werden bei den partitionierenden Clusterverfahren unterschiedliche Definitionen des Abstandes von Objekten verwendet. Im Gegensatz zu hierarchischen Verfahren ist jedoch die Zuordnung von Fällen zu Clustern revidierbar, d.h. im Prozess des Clusterns können Fälle erneut einem anderen Clustern zugeordnet werden.

Im Folgenden wird das von MacQueen entwickelte *k-Means* Verfahren dargestellt, als dem im Bereich der Sozialwissenschaften am häufigsten verwendeten nicht-hierarchischen Clusterverfahren (vgl. Bortz 1999). Das *k-Means* Verfahren bezieht sich auf die *Schwerpunkte* bzw. Mittelpunkte einer definierten Anzahl von Clustern (k), was auch im Namen der Methode zum Ausdruck kommt.

Das Vorgehen besteht aus folgenden iterativen Schritten:

- „Man erzeugt eine Anfangspartition mit k Clustern.
- Beginnend mit dem 1. Objekt im 1. Cluster werden für alle Objekte die euklidischen Distanzen zu allen Clusterschwerpunkten [...] bestimmt.
- Trifft man auf ein Objekt, das zu dem Schwerpunkt des eigenen Clusters eine größere Distanz aufweist als zum Schwerpunkt eines anderen Clusters, wird dieses Objekt in dieses Cluster verschoben.
- Die Schwerpunkte der beiden durch diese Verschiebung veränderten Cluster werden neu berechnet.
- Man wiederholt Schritt 2 bis Schritt 4, bis sich jedes Objekt in einem Cluster befindet, zu dessen Schwerpunkt es im Vergleich zu den übrigen Clustern die geringste Distanz aufweist“ (Bortz 1999: 560).

Als Ausgangspartition bzw. Startgruppierung des *k-Means* Verfahrens dient häufig das Ergebnis einer hierarchischen Clusteranalyse. Ziel ist es, dieses Ergebnis zu optimieren. Im Rahmen der Navigationsanalyse kann das *k-Means* Verfahren jedoch nicht zur Validierung und Optimierung genutzt werden. Die Durchführung einer *k-Means* Clusteranalyse im Anschluss an eine hierarchische Clusteranalyse erfordert als Voraussetzung eine Datenmatrix. Diese ist jedoch bei den vorliegenden Daten nicht vorhanden. Vorhanden ist lediglich die Distanzmatrix, bestehend aus den Distanzen der paarweisen Vergleiche aller Sequenzen. Darüber hinaus gehende Variablen stehen im vorliegenden empirischen Datensatz nicht zur Verfügung. Das *k-Means* Verfahren ist auf der Basis einer Distanzmatrix nicht anwendbar, da der Schwerpunkt („centroid“) für jedes Cluster in einem n-dimensionalen Raum berechnet werden muss. Auf dieser Grundlage wird die Zuordnung der Fälle in die entsprechenden Cluster mit dem Ziel der Minimierung der euklidischen Summe der Quadrate („Euclidean Sum of Squares“, ESS) vorgenommen. Allein auf Grundlage der Distanzmatrix können jedoch keine Schwerpunkte („centroids“) berechnet werden.[79]

Als Alternative zur Optimierung der Clusterlösung hierarchischer Clusterverfahren schlägt Wishart[80] die Verwendung der in ClustanGraphics implementierten *Bootstrap*-Validierung vor, bei der überprüft wird, ob die Clusterlösungen signifikant sind im Vergleich zu einer zufälligen Ausgangsverteilung. Darüber hinaus schlägt Wishart die Verwendung des in ClustanGraphics implementierten Verfahrens des *Multidimensional Scaling* (MDS) vor, das an der Abstandsmatrix ansetzt. Das Ergebnis ist ein Set neu berechneter Variablen, das einer räumlichen Anordnung der Abstände entspricht. Mit diesen Daten als Ausgangspunkt kann mit Hilfe der räumlichen Darstellung in Form von Scatterplots der Zusammenhang der Sequenzen veranschaulicht werden und darüber hinaus bieten diese Daten die Möglichkeit der Verwendung einer *k-Means* Analyse.

79 Die fehlende Möglichkeit zur Berechnung von Schwerpunkten („centroids“) ist auch der Grund für die Nichtanwendbarkeit von Clusterverfahren, die auf der Berechnung von Schwerpunkten beruhen (s. Kap. 9.3).

80 Private Email-Korrespondenz vom 24.05.06.

9.2.3 Dendogramm

Das konkrete Vorgehen der hierarchischen Clusteranalyse kann grafisch anhand eines *Dendogramms (Graph Tree)* veranschaulicht werden.[81] Die schrittweise Vereinigung der jeweils zwei ähnlichsten Gruppen wird dabei als grafische Baumstruktur dargestellt und enthält damit Informationen über den Aggregationsprozess, d.h. darüber, in welcher Abfolge die Sequenzen des Datensatzes fusioniert werden.

Cluster sind im Dendogramm durch senkrechte Linien gekennzeichnet, waagerechte Linien verweisen auf die Objekte des Clusters. Zusammengehörende waagerechte und senkrechte Linien können somit als Klammer für Cluster und deren Objekte gelesen werden. Dendogramme stellen den Prozess der Fusionierung von Fällen dar und verdeutlichen damit deren hierarchischen Aufbau: es wird erkennbar, welche Objekte zu welchen Clustern und welche Cluster zu welchen neuen Clustern fusioniert werden. Darüber hinaus wird durch die Länge der waagerechten Linie der Abstand unterschiedlicher Cluster zueinander verdeutlicht, sowie die Reihenfolge der Fusionierung.

Zusätzlich wird die *Monotonieeigenschaft* von hierarchischen Clusteranalysen verdeutlicht: Die Abstände bzw. Fusionswerte der Cluster steigen im Verlauf der Aggregation an. In den ersten Schritten der hierarchischen Clusteranalyse zeigt das Dendogramm eine große Anzahl von Clustern, die in sich sehr homogen sind. Im Verlauf des Clusterprozesses werden immer heterogenere Cluster fusioniert bis schließlich alle Objekte in einem Cluster zusammengefasst werden. Generell sind Dendogramme wichtige Hilfsmittel zur Festlegung der für die Analyse geeigneten Anzahl der Cluster. Neben der Verwendung des Elbow-Kriteriums ist vor allem die *Interpretierbarkeit* der resultierenden Clusterlösung in der Regel das Hauptkriterium zur Festlegung der Anzahl der Cluster (vgl. Brüderl / Scherer 2005: 6).

9.3 Formale Beschreibung der Clusteralgorithmen

In diesem Kapitel wird die Entscheidung über das im Rahmen der Navigationsanalyse verwendete clusteranalytische Verfahren begründet. Dazu werden die in TDA implementierten Verfahren der Clusteranalyse zunächst auf *formaler Ebene* beschrieben und anschließend hinsichtlich ihrer Ergebnisse inhaltlich analysiert und interpretiert. Die empirische Datenbasis der Analyse und des Vergleichs in diesem Kapitel ist der Teil des Gesamtdatensatzes (vgl. Kap.10.1; *Datenerhebung und Datenbasis*), der sich auf die Prozesse der *Mikronavigation* innerhalb der *Lerneinheit „Maße der zentralen Tendenz“* (513) bezieht: es handelt sich dabei um 475 Sequenzen mit insgesamt 1542 Einheiten.

81 Beispiele für Dendogramme befinden sich in Kap. 7, *Sequenzanalyse am Beispiel*.

Innerhalb von TDA werden hierarchische Clusteranalysen mit der Befehls-Syntax *hcls* und dem Parameter *opt* umgesetzt (vgl. Abb. 29):[82] single-link (1), complete-link (2), weighted average (3), weighted centroid (4), group average (5), unweighted centroid (6), ward's minimum variance (7).

Von diesen sieben in TDA implementierten Clusterverfahren sind die Algorithmen von drei Verfahren nicht auf die vorliegenden Daten anwendbar:[83] Daher werden im Folgenden die Clusterlösungen der Verfahren

- complete-link (2),
- weighted-average (3),
- group-average (5) sowie
- ward's minimum variance (7)

analysiert und miteinander verglichen. Im Fokus stehen dabei die Unterschiede und Gemeinsamkeiten der Clusterlösungen in Form der Zuordnung von Fällen (Sequenzen) zu spezifischen Clustern.

```
hcls (
   opt=...,      option, def. 1
                 1 = single-link
                 2 = complete-link
                 3 = weighted average
                 4 = weighted centroid
                 5 = group average
                 6 = unweighted centroid
                 7 = Ward's minimum variance
```

Abbildung 29: In TDA implementierte clusteranalytische Verfahren (TDA-Manual 2005: 923)

Den gemeinsamen Ausgangspunkt des Vergleichs bildet eine Distanzmatrix als Ergebnis einer Optimal-Matching Analyse:[84] Die jeweiligen Clusterverfahren setzen an dieser Distanzmatrix mit unterschiedlichen Clusteralgorithmen an. Zur differenzierten Analyse wird auf die Clusterlösung mit 10 und 28 Clustern zurückgegriffen (vgl. Kap. 9.4, *Begründung der Wahl der Clusteranzahl*). Für jedes verwendete Clusterverfahren wird die jeweilige *Struktur des Agglomerationsprozesses* anhand eines Dendogramms verdeutlicht, sowie die Anzahl der Sequenzen in den einzelnen Clustern tabellarisch dargestellt.

82 Die Ziffern in Klammern der folgenden Aufzählung beziehen sich auf den *opt*-Parameter des *hcls*-Befehls in TDA.

83 Für die vorliegenden Daten nicht anwendbar sind die Clusterverfahren s*ingle-link* (1), *weighted centroid* (4) und *unweighted centroid (6)*. So führt beispielsweise das single-link Clusterverfahren bei den vorliegenden Daten zu kettenförmigen Clustergebilden (Chaining-Effekte). Bortz (1999: 554) bewertet auf Grund dieses Effektes das single-linkage Clusterverfahren als für sozialwissenschaftliche Zwecke wenig geeignet. Daneben sind Verfahren, die auf der Berechnung des Clusterschwerpunktes („centroid“) beruhen nicht verwendbar, da Clusterschwerpunkte auf Grundlage einer Distanzmatrix nicht errechnet werden können.

84 Die Optimal-Matching Analyse wird dabei auf Grundlage der default-Definition der Substitutionskosten durchgeführt, vgl. Kap. 9.1.

9.3.1 Clusteralgorithmus „complete link“

Das folgende Dendogramm (vgl. Abb. 30) zeigt die Struktur des Agglomerationsprozesses des „complete link“-Clusteralgorithmus. Ausgangspunkt ist die Clusterlösung mit 28 Clustern.

Die Häufigkeitsverteilung dieser Lösung mit 28 Clustern ist in der folgenden Tabelle (vgl. Abb. 31) dokumentiert. Auffällig an dieser Clusterlösung ist die stark ungleiche Zuordnung von Fällen zu Clustern. So enthalten die beiden zahlenmäßig größten Cluster (1, 3) insgesamt 75,1 % aller Fälle. Das größte Cluster (1) enthält dabei 54,5 % aller Fälle. Im Gegensatz zu diesen beiden großen Clustern fallen die Cluster auf, die lediglich aus einer geringen Anzahl von Fällen bzw. einem einzigen Fall bestehen (7, 8, 9, 13, 18, 19, 20, 21, 24, 25, 27, 28). Dies sind insgesamt 12 Cluster.

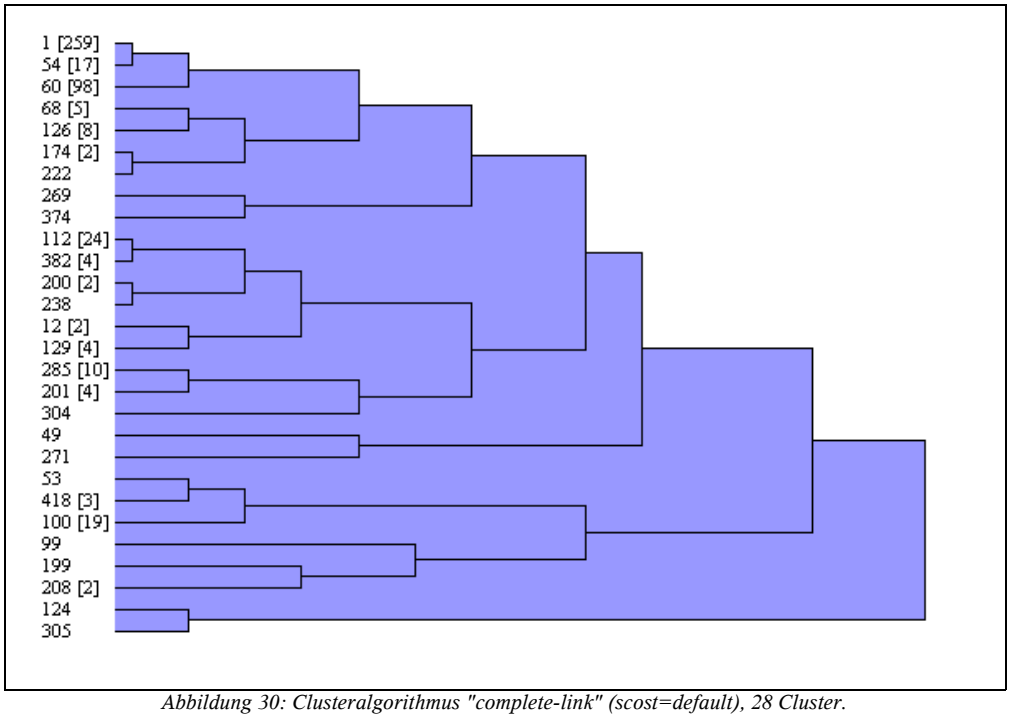

Abbildung 30: Clusteralgorithmus "complete-link" (scost=default), 28 Cluster.

Die Häufigkeitsverteilung der Lösung mit 10 Clustern (vgl. Abb. 32) bestätigt die oben beschriebene Tendenz. Auffällig ist auch hier die stark ungleiche Verteilung von Fällen zu Clustern. Die beiden größten Cluster (1, 3) enthalten bei dieser Lösung 89,9% aller Fälle; das größte Cluster (1) enthält 82,1 %. Im Gegensatz dazu stehen 6 Cluster mit weniger als 5 Fällen (2, 5, 6, 8, 9, 10), sowie 2 Cluster (5, 8) mit lediglich einem Fall.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	259	54,5	54,5	54,5
	2	17	3,6	3,6	58,1
	3	98	20,6	20,6	78,7
	4	5	1,1	1,1	79,8
	5	8	1,7	1,7	81,5
	6	2	,4	,4	81,9
	7	1	,2	,2	82,1
	8	1	,2	,2	82,3
	9	1	,2	,2	82,5
	10	24	5,1	5,1	87,6
	11	4	,8	,8	88,4
	12	2	,4	,4	88,8
	13	1	,2	,2	89,1
	14	2	,4	,4	89,5
	15	4	,8	,8	90,3
	16	10	2,1	2,1	92,4
	17	4	,8	,8	93,3
	18	1	,2	,2	93,5
	19	1	,2	,2	93,7
	20	1	,2	,2	93,9
	21	1	,2	,2	94,1
	22	3	,6	,6	94,7
	23	19	4,0	4,0	98,7
	24	1	,2	,2	98,9
	25	1	,2	,2	99,2
	26	2	,4	,4	99,6
	27	1	,2	,2	99,8
	28	1	,2	,2	100,0
	Gesamt	475	100,0	100,0	

Abbildung 31: Häufigkeiten des Clusteralgorithmus "complete-link" (scost=default), 28 Cluster.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	390	82,1	82,1	82,1
	2	2	,4	,4	82,5
	3	37	7,8	7,8	90,3
	4	14	2,9	2,9	93,3
	5	1	,2	,2	93,5
	6	2	,4	,4	93,9
	7	23	4,8	4,8	98,7
	8	1	,2	,2	98,9
	9	3	,6	,6	99,6
	10	2	,4	,4	100,0
	Gesamt	475	100,0	100,0	

Abbildung 32: Häufigkeiten des Clusteralgorithmus "complete-link" (scost=default), 10 Cluster.

9.3.2 Clusteralgorithmus „weighted average“

Das folgende Dendogramm (vgl. Abb. 33) zeigt die Struktur des Agglomerationsprozesses des „weighted average“-Clusterverfahrens. Ausgangspunkt ist die Clusterlösung mit 28 Clustern.

Die Häufigkeitsverteilung dieser Lösung mit 28 Clustern ist in der folgenden Tabelle (vgl. Abb. 34) dokumentiert. Auch an dieser Clusterlösung ist die stark ungleiche Verteilung von Fällen zu Clustern auffällig. Die beiden größten Cluster (1, 20) enthalten insgesamt 82,3% aller Fälle. Das größte Cluster (1) enthält da-

bei 70,9% aller Fälle. Im Vergleich dazu enthalten die Cluster 8, 10, 13, 14, 15, 16, 17, 18, 21, 22, 23, 25, 26, 27, 28 enthalten jeweils einen Fall (insgesamt 15 Cluster).

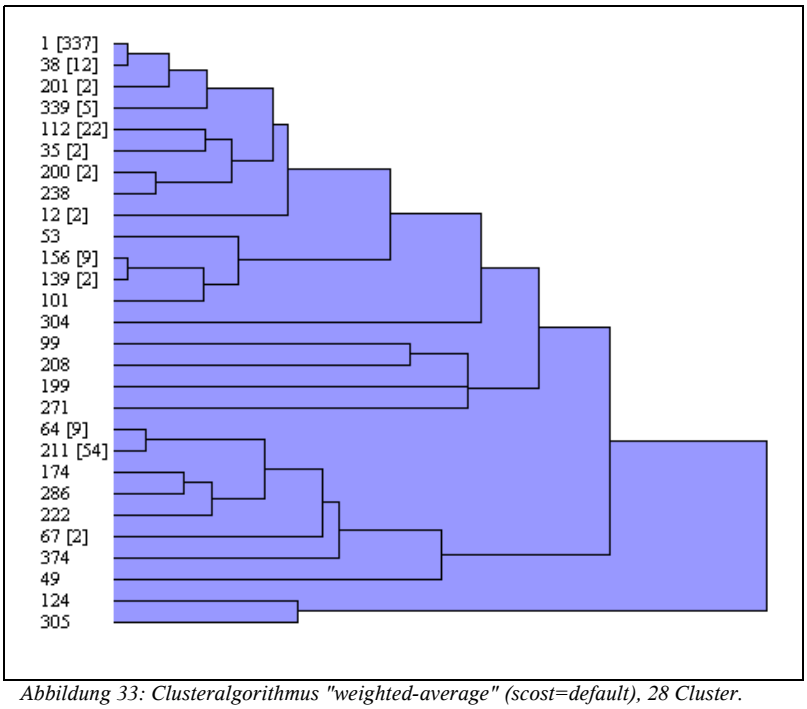

Abbildung 33: Clusteralgorithmus "weighted-average" (scost=default), 28 Cluster.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	337	70,9	70,9	70,9
	2	12	2,5	2,5	73,5
	3	2	,4	,4	73,9
	4	5	1,1	1,1	74,9
	5	22	4,6	4,6	79,6
	6	2	,4	,4	80,0
	7	2	,4	,4	80,4
	8	1	,2	,2	80,6
	9	2	,4	,4	81,1
	10	1	,2	,2	81,3
	11	9	1,9	1,9	83,2
	12	2	,4	,4	83,6
	13	1	,2	,2	83,8
	14	1	,2	,2	84,0
	15	1	,2	,2	84,2
	16	1	,2	,2	84,4
	17	1	,2	,2	84,6
	18	1	,2	,2	84,8
	19	9	1,9	1,9	86,7
	20	54	11,4	11,4	98,1
	21	1	,2	,2	98,3
	22	1	,2	,2	98,5
	23	1	,2	,2	98,7
	24	2	,4	,4	99,2
	25	1	,2	,2	99,4
	26	1	,2	,2	99,6
	27	1	,2	,2	99,8
	28	1	,2	,2	100,0
	Gesamt	475	100,0	100,0	

Abbildung 34: Häufigkeiten des Clusteralgorithmus "weighted-average" (scost=default), 28 Cluster.

Die Häufigkeitsverteilung der Lösung mit 10 Clustern (Abb. 35) bestätigt die oben beschriebene Tendenz für die Lösung mit 28 Clustern. Die beiden größten Cluster (1, 63) enthalten bei dieser Lösung 97,7% aller Fälle; das größte Cluster (1) enthält 67,6 %. Alle anderen Cluster enthalten weniger als 4 Fälle, wobei die Cluster 2, 3, 4, 5, 8, 10 lediglich einen Fall enthalten.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	321	67,6	67,6	67,6
	2	1	,2	,2	67,8
	3	1	,2	,2	68,0
	4	1	,2	,2	68,2
	5	1	,2	,2	68,4
	6	143	30,1	30,1	98,5
	7	3	,6	,6	99,2
	8	1	,2	,2	99,4
	9	2	,4	,4	99,8
	10	1	,2	,2	100,0
	Gesamt	475	100,0	100,0	

Abbildung 35: Häufigkeiten des Clusteralgorithmus "weighted-average" (scost=default), 10 Cluster.

9.3.3 Clusteralgorithmus „group-average“

Das folgende Dendogramm (vgl. Abb. 36) zeigt die Struktur des Agglomerationsprozesses des „group-average“-Clusterverfahrens. Ausgangspunkt ist die Clusterlösung mit 28 Clustern.

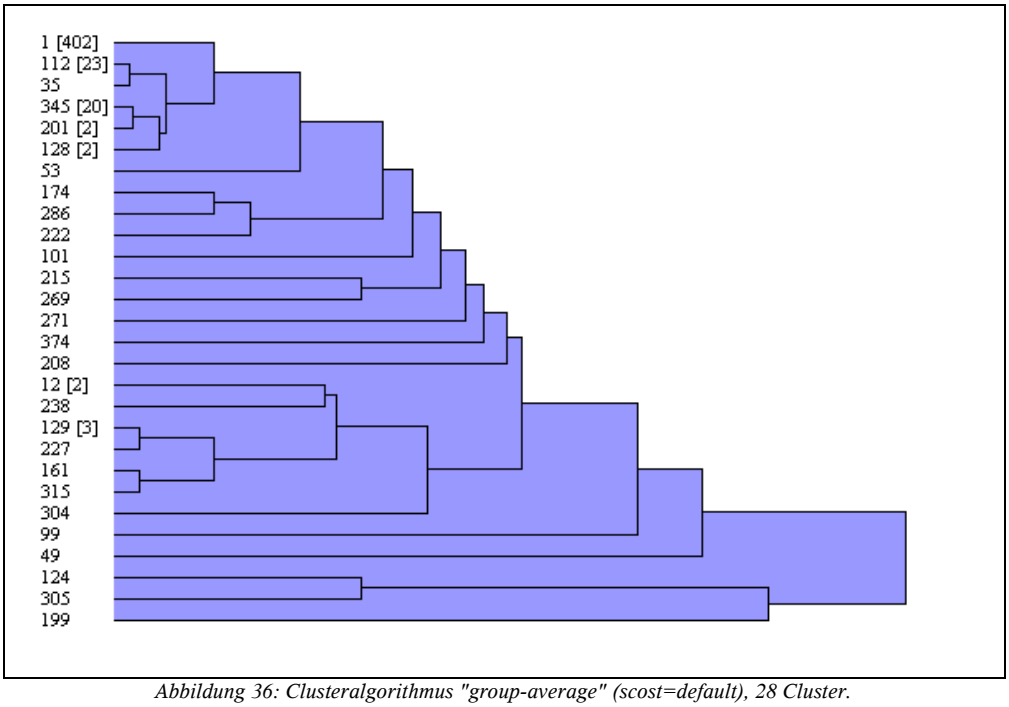

Abbildung 36: Clusteralgorithmus "group-average" (scost=default), 28 Cluster.

In Abbildung 38 sind die Häufigkeitsverteilungen der Lösung mit 28 Clustern dargestellt. Die beiden größten Cluster (1, 2) enthalten insgesamt 89% aller Fälle; das größte Cluster (1) enthält 84,6% aller Fälle. Auch hier ist die große Anzahl von Clustern auffällig, die lediglich einen Fall enthalten (3, 7, 8, 9, 10, 12, 13, 14, 15, 16, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28).

Die Häufigkeitsverteilung der Lösung mit 10 Clustern (vgl. Abb. 37) zeigt, dass 97,9% aller Fälle in den beiden größten Clustern (1, 3) enthalten sind; das größte Cluster (1) deckt in dieser Clusterlösung 95,8% aller Fälle ab.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	455	95,8	95,8	95,8
	2	2	,4	,4	96,2
	3	1	,2	,2	96,4
	4	1	,2	,2	96,6
	5	1	,2	,2	96,8
	6	10	2,1	2,1	98,9
	7	1	,2	,2	99,2
	8	1	,2	,2	99,4
	9	2	,4	,4	99,8
	10	1	,2	,2	100,0
	Gesamt	475	100,0	100,0	

Abbildung 37:Häufigkeiten des Clusteralgorithmus "group-average" (scost=default), 10 Cluster.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	402	84,6	84,6	84,6
	2	21	4,4	4,4	89,1
	3	1	,2	,2	89,3
	4	20	4,2	4,2	93,5
	5	3	,6	,6	94,1
	6	2	,4	,4	94,5
	7	1	,2	,2	94,7
	8	1	,2	,2	94,9
	9	1	,2	,2	95,2
	10	1	,2	,2	95,4
	11	2	,4	,4	95,8
	12	1	,2	,2	96,0
	13	1	,2	,2	96,2
	14	1	,2	,2	96,4
	15	1	,2	,2	96,6
	16	1	,2	,2	96,8
	17	2	,4	,4	97,3
	18	1	,2	,2	97,5
	19	3	,6	,6	98,1
	20	1	,2	,2	98,3
	21	1	,2	,2	98,5
	22	1	,2	,2	98,7
	23	1	,2	,2	98,9
	24	1	,2	,2	99,2
	25	1	,2	,2	99,4
	26	1	,2	,2	99,6
	27	1	,2	,2	99,8
	28	1	,2	,2	100,0
	Gesamt	475	100,0	100,0	

Abbildung 38: Häufigkeiten des Clusteralgorithmus "group-average" (scost=default), 28 Cluster.

9.3.4 Clusteralgorithmus „ward's minimum variance“ (Ward)

Das folgende Dendogramm (vgl. Abb. 39) zeigt die Struktur des Agglomerationsprozesses des Ward-Clusterverfahrens. Ausgangspunkt ist die Clusterlösung mit 28 Clustern.

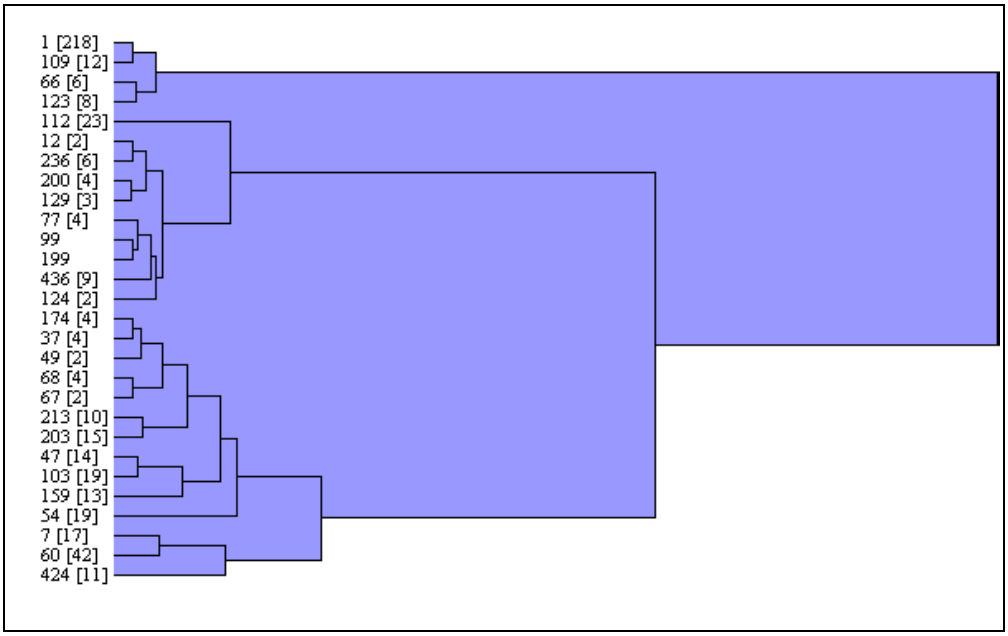

Abbildung 39: Clusteralgorithmus "ward's minimum variance" (scost=default), 28 Cluster.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	218	45,9	45,9	45,9
	2	12	2,5	2,5	48,4
	3	6	1,3	1,3	49,7
	4	8	1,7	1,7	51,4
	5	23	4,8	4,8	56,2
	6	2	,4	,4	56,6
	7	6	1,3	1,3	57,9
	8	4	,8	,8	58,7
	9	3	,6	,6	59,4
	10	4	,8	,8	60,2
	11	1	,2	,2	60,4
	12	1	,2	,2	60,6
	13	9	1,9	1,9	62,5
	14	2	,4	,4	62,9
	15	4	,8	,8	63,8
	16	4	,8	,8	64,6
	17	2	,4	,4	65,1
	18	4	,8	,8	65,9
	19	2	,4	,4	66,3
	20	10	2,1	2,1	68,4
	21	15	3,2	3,2	71,6
	22	14	2,9	2,9	74,5
	23	19	4,0	4,0	78,5
	24	13	2,7	2,7	81,3
	25	19	4,0	4,0	85,3
	26	17	3,6	3,6	88,8
	27	42	8,8	8,8	97,7
	28	11	2,3	2,3	100,0
	Gesamt	475	100,0	100,0	

Abbildung 40: Häufigkeiten des Clusteralgorithmus "ward's minimum variance" (scost=default), 28 Cluster.

Abbildung 39 zeigt die Häufigkeitsverteilung der Lösung mit 28 Clustern. Die beiden größten Cluster (1, 27) enthalten insgesamt 54,7% aller Fälle; das größte Cluster (1) enthält 45,9% aller Fälle. Insgesamt enthalten 2 Cluster lediglich einen Fall (11, 12).

Die entsprechende Lösung mit 10 Clustern (vgl. Abb. 41) zeigt, dass die beiden größten Cluster (1, 9) insgesamt 63,8% aller Fälle enthalten, wobei im größten Cluster (1) 51,4% aller Fälle enthalten sind. Bei dieser Clusterlösung gibt es keine Cluster mit lediglich einen Fall.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	244	51,4	51,4	51,4
	2	23	4,8	4,8	56,2
	3	32	6,7	6,7	62,9
	4	16	3,4	3,4	66,3
	5	25	5,3	5,3	71,6
	6	33	6,9	6,9	78,5
	7	13	2,7	2,7	81,3
	8	19	4,0	4,0	85,3
	9	59	12,4	12,4	97,7
	10	11	2,3	2,3	100,0
	Gesamt	475	100,0	100,0	

Abbildung 41: Häufigkeiten des Clusteralgorithmus "ward's minimum variance" (scost=default), 10 Cluster.

9.3.5 Begründung der Wahl des Clusteralgorithmus

Aus der Darstellung der *formalen* Analyse der Clusteralgorithmen wird für das weitere Vorgehen der Navigationsanalyse die Verwendung der Clusterverfahren „complete-link“, „weighted-average“ sowie „group-average“ verworfen.

Die Lösungen der Clusterverfahren „complete-link“, „weighted-average“ sowie „group-average“ sind auf der Ebene von 10 wie auch von 28 Clustern zu *wenig differenziert* in Hinblick auf die Zuordnung der Fälle zu Clustern. Die Fälle sind stark ungleich verteilt: Einzelne große Cluster dominieren die Clusterlösung; diesen sehr großen Clustern stehen Cluster mit sehr wenigen bzw. einzelnen Fällen gegenüber.

Die oben dargestellten *formalen* Kennzeichen der Clusterlösungen dienen daher als Ausgangspunkt für eine *inhaltliche* Begründung der Wahl des Clusterverfahrens. Allgemein kann festgehalten werden, dass die genannten Clusteralgorithmen nicht in der Lage sind, die vorliegenden Sequenzdaten angemessen zu gruppieren, d.h. eine angemessene *inhaltliche* Interpretation der Cluster zu ermöglichen. So ist einerseits eine konsistente inhaltliche Interpretation der beschriebenen größten Cluster nicht möglich; die Clusterlösungen stellen sich als *heterogene* Gruppen unterschiedlicher Fälle dar, die nicht zusammenfassend beschrieben werden

können. Auf der anderen Seite stehen Cluster mit lediglich einem Fall, die sich nur sehr schwer von den zahlenmäßig großen Clustern abgrenzen lassen und inhaltlich wenig Aussagekraft besitzen.

Im Rahmen der vorliegenden Navigationsanalyse wird daher das Ward-Clusterverfahren verwendet: In *formaler* Hinsicht stellen die Clusterlösungen auf der Ebene von 10 wie auch von 28 Clustern eine deutlich *differenzierte* Gruppierung gegenüber den verglichenen Clusterlösungen dar. Dies zeigt sich sowohl in der Abdeckung durch das zahlenmäßig größte Cluster als auch in der Zuordnung von Fällen zu den kleineren Clustern.

Auch in *inhaltlicher* Hinsicht ist die Clusterlösung des Ward-Verfahrens als am differenziertesten zu beurteilen. Dies zeigt sich beispielhaft bei der Clusterlösung mit 28 Clustern, und dort speziell bei dem größten Cluster (1): Dieses Cluster ist inhaltlich (und formal) sehr homogen und besteht ausschließlich aus einmaligen Zugriffen auf die Wissenseinheit *Orientierung* (2642). Die Clusterlösungen der verworfenen Clusterverfahren hinsichtlich der zahlenmäßig großen Cluster ist wesentlich undifferenzierter. In ihnen zeigt sich eine Heterogenität von Fällen; neben einmaligen Aufrufen sind vielfältige weitere Sequenzen (Fälle) diesen Clustern zugeordnet (vgl. Kap. 9.1.4).

cg_subcostdefault_ward_28c * cg_completelink_28c Kreuztabelle

Anzahl

		cg_completelink_28c																											Gesamt
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
cg_subcostdefault_ward_28c	1	218	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	218
	2	2	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	12
	3	2	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	8
	5	0	0	1	0	1	0	0	0	0	18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	23
	6	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	7	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
	8	0	0	0	0	1	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
	10	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4
	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
	13	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5	3	0	0	0	0	0	0	0	0	0	0	0	0	9
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2
	15	0	0	1	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	16	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2
	18	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	19	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	20	0	0	1	0	5	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	10
	21	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	15
	22	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
	23	16	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19
	24	3	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	13
	25	1	17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19
	26	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17
	27	0	0	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42
	28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	6	0	0	0	0	0	11
Gesamt		259	17	98	5	8	2	1	1	1	24	4	2	1	2	4	10	4	1	1	1	1	3	19	1	1	2	1	1	475

Abbildung 42: Kreuztabelle Clusterlösung ward und complete-link (28 Cluster).

Beispielhaft kann dieser höhere Grad der inhaltlichen Differenzierung des Ward-Verfahrens anhand eines Vergleichs mit dem „complete link"-Verfahren dargestellt werden.[85] Dazu werden zunächst die jeweiligen Clusterlösungen mit Hilfe einer Kreuztabelle (vgl. Abb. 42) gegenübergestellt („complete-link": 1. Zeile;

85 Dieser höhere Grad inhaltlicher Differenzierung des Ward-Verfahrens findet sich auch im Vergleich mit den Clusterverfahren „weighted-average" und „group-average". Die entsprechenden Kreuztabellen verdeutlichen, dass die Clusterlösungen für das größte Cluster jeweils weniger differenziert und inhaltlich heterogener sind als die Clusterlösung des Ward-Verfahrens.

„ward“: 1. Spalte). Analysiert wird die Zusammensetzung des Clusters 1 in der Ward-Clusterlösung mit dem Cluster 1 der „complete-link“-Clusterlösung: Es wird deutlich, dass das Cluster 1 („ward“) vollständig im Cluster 1 („complete-link“) enthalten ist.

Bei der inhaltlichen Analyse wurde bereits darauf hingewiesen, dass das Cluster 1 („ward“) sehr homogen ist (vgl. Kap. 9.1.4, *Inhaltliche Analyse der unterschiedlichen Clusterlösungen*). Das „complete-link“-Verfahren fusioniert in das Cluster 1 noch weitere 41 Fälle. Das Ward-Verfahrens schlägt für diese Fälle eine davon abweichende Clusterzuordnung vor. Dabei sind die Cluster 22 (vgl. Abb. 43) und 23 (vgl. Abb. 44) besonders aussagekräftig, da diese in der Clusterlösung („ward“) eigenständige Cluster darstellen. Am konkreten Beispiel der Zuordnung von Fällen zu diesen Clustern ist der höhere Grad der Differenzierung und daraus resultierend der Homogenität der Ward-Clusterlösung deutlich ablesbar: Während das „complete-link“-Verfahren *ein* Cluster vorschlägt, schlägt das Ward-Verfahren die Zuordnung der Fälle zu *drei unterschiedlichen* Clustern (1, 22 und 23) vor.

Aufgrund dieses höheren *formalen* Differenzierungsgrades des „ward“-Clusterverfahrens ist dessen Clusterlösung auch *inhaltlich homogener* als die Clusterlösung des „complete-link“-Verfahrens.

2642	3597	.	.
2642	3597	.	.
2642	3597	.	.
2642	3597	.	.
2642	3597	.	.
2642	3597	.	.
2642	3597	.	.
2642	3597	.	.
2642	3597	.	.
2642	3597	.	.
2642	3597	2896	.
2643	3597	2896	.
3597	2642	3597	.
3597	2642	3597	2896

Abbildung 43: Cluster 22 ("ward")

2642	3597	2642	.	.	.	.
2642	3597	2642	.	.	.	.
2642	3597	2642	.	.	.	.
2642	3597	2642	.	.	.	.
2642	3597	2642	2642	.	.	.
2642	3597	2896	3597	2642	.	.
2642	3597	2896	3597	2642	3597	.
2642	3597	2900	2900	2642	2642	.
3597	2642	.	.	.	.	.
3597	2642	.	.	.	.	.
3597	2642	.	.	.	.	.
3597	2642	.	.	.	.	.
3597	2642	.	.	.	.	.
3597	2642	.	.	.	.	.
3597	2642	2642	3597	2642	.	.
3597	3597	2642	3597	2642	2642	2642

Abbildung 44: Cluster 23 ("ward")

9.4 Begründung der Wahl der Clusteranzahl

Ziel der Clusteranalyse ist die Gruppierung von Fällen in möglichst homogene Cluster, wobei die Cluster untereinander möglichst heterogen sind (vgl. 9.2: 108). Die Wahl einer bestimmten Anzahl an Clustern stellt neben der Wahl des Clusterverfahrens eine der zentralen Entscheidung bei der Clusteranalyse dar (vgl. Fraley / Raftery 1998, Bortz 1999; Backhaus / Erichson / Plinke / Wulff 2000). In ihr kommt die Frage nach der in den Daten enthaltenen Struktur zum Ausdruck.

Für die Ermittlung der *idealen* Clusteranzahl eines Datensatzes existieren unterschiedliche Kriterien und Vorgehensweisen, z.B. das Ansteigen der Fehlerquadratsumme (F'usionskoeffizient) beim Ward-Verfahren oder die Anwendung einer t-Teststatistik. Jedoch weist Micheel (2002: 52) in aller Deutlichkeit darauf hin, dass diese Kriterien lediglich aus *mathematischer* Perspektive zu *optimalen Lösungen* führen. Zudem besteht diese *optimale Lösung* in der Regel in einer geringen Anzahl von Clustern, die für sozialwissenschaftlichen Fragestellungen jedoch meist nicht angemessen ist: Die *optimale mathematische Lösung* besteht für das Ward-Verfahren aufgrund der Zunahme der Fehlerquadratsumme in einer 2-Cluster-Lösung, d.h. für den letzten Fusionsschritt ist der Fusionskoeffizient am höchsten. So beträgt der t-Wert für die 2-Cluster-Lösung für die analysierten Daten dieses Kapitels 377,10. Statistisch signifikant auf einem 5% Niveau sind jedoch alle Clusterlösungen mit weniger als 56 Clustern (vgl. Abb. 45).

Daher schlägt Micheel (2002) vor, diese mathematischen Kriterien lediglich als Anhaltspunkt für die Entscheidung über die Clusteranzahl zu verwenden. Die endgültige Entscheidung für eine Clusterlösung kann dann nur *inhaltlich* in Bezug zur übergeordneten Fragestellung getroffen werden.[86] Auch Erzberger (2001: 153) interpretiert das Ansteigen des Fusionskoeffizienten bei der Ward-Clusteranalyse lediglich als Hinweis auf den „Entscheidungsraum“, in dem die den Daten ädaquate Clusterlösung zu finden ist.

In der hier vorliegenden Navigationsanalyse wird aufgrund der heuristisch-explorativen Zielrichtung die Wahl der Clusteranzahl *inhaltlich-pragmatisch* getroffen: Aufgrund einer umfangreichen *formalen* und *inhaltlichen* Analyse unterschiedlicher Clusterlösungen wird einerseits eine Lösung mit 10 Clustern verwendet, um einen allgemeinen Überblick über die Grundstruktur der analysierten Fälle zu erhalten.[87] Vor allem bei der Interpretation von Clustern mit hoher Fallzahl, die aus inhaltlicher Perspektive eine weitergehende Differenzierung nahe legen, wird auf eine Lösung mit 28 Clustern zurückgegriffen.[88] Aus inhaltlicher Pers-

86 Darüber hinaus schlägt Micheel (2002) die Verwendung clusterspezifischer Kenngrößen für die unterschiedlichen Clusterlösungen vor, wie z.B. Mittelwerte und Streuungen von Variablen. Dieses Vorgehen ist im Rahmen der hier vorgestellten Navigationsanalyse nicht möglich, da die analysierten Daten sich auf eine Distanzmatrix beziehen und darüber hinaus keine Variablen für die Berechnung zur Verfügung stehen.

87 Für die in diesem Kapitel analysierten Daten (s.o.) beträgt der t-Wert für die 10-Cluster-Lösung 45,43 und ist damit statistisch signifikant auf einem 5% Niveau.

88 Für die in diesem Kapitel analysierten Daten (s.o.) beträgt der t-Wert für die 28-Cluster-Lösung 8,04 und ist damit statistisch signifikant auf einem 5% Niveau.

pektive zeigt sich beispielsweise bei der 28-Cluster-Lösung die homogenste Fusionierung des Clusters mit nur einem Zugriff auf die untersuchte Lerneinheit (513).[89]

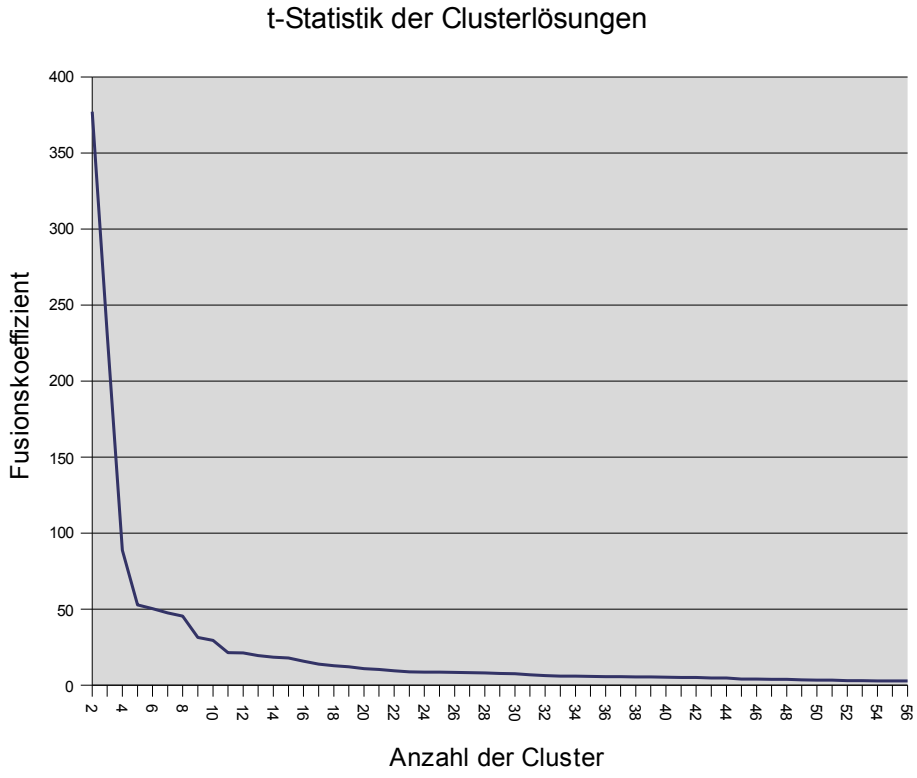

Abbildung 45: t-Statistik der Clusterlösungen (Fusionskoeffizienten).

89 für die formal-inhaltliche Interpretation und Angemessenheit dieser Entscheidung, vgl. Kap. 11, *Ergebnisse der Navigationsanalyse*.

10 Durchführung der Navigationsanalyse

In diesem Kapitel wird die konkrete Durchführung der Navigationsanalyse mittels Optimal-Matching dargestellt.

Im einleitenden Kapitel werden die verwendete *Methode der Datenerhebung* sowie die *empirische Datenbasis* der Navigationsanalyse beschrieben. Aufbauend auf den Ausführungen in den vorangehenden Kapiteln werden die grundlegenden *konzeptionellen* und *theoretischen* Entscheidungen der Durchführung zusammenfassend dargestellt. Daran anschließend wird die konkrete Durchführung auf *programmtechnischer Ebene* auf der Grundlage der verwendeten TDA-Syntax beschrieben.

In folgenden Kapitel 11 werden dann die Ergebnisse der durchgeführten Analyse detailliert ausgeführt.

10.1 Datenerhebung und Datenbasis

Die empirische Grundlage der durchgeführten Navigationsanalyse beruht auf der Analyse von *Verhaltensspuren* in Form von Logdaten. Neben Befragung und Beobachtung versteht Diekmann (vgl. 2007: 629) Verhaltensspuren als eine eigenständige Methode der Datenerhebung (vgl. auch Verfahren der Verhaltensdokumentation, Kap. 2.1.3, *Bilddaten*).

Verhaltensspuren gehören zu den nicht-reaktiven, unaufdringlichen Erhebungsverfahren („unobtrusive measures“, vgl. Webb et al. 1966) wie z.B. auch die verdeckte Beobachtung. Die Erhebung von Verhaltensspuren in Form von Logdaten greift nicht aktiv in das soziale Geschehen ein, vielmehr werden die Logdaten quasi nebenbei aufgezeichnet. Das Produzieren von Logdaten ist weder die zentrale Intention der Nutzung der Online-Plattform noch werden sie vom Nutzenden speziell für die wissenschaftliche Analyse erzeugt.

So versteht auch Priemer (2004) die Protokollierung von Logdaten in Online-Umgebungen als unbemerkte und detailgenaue objektive Aufzeichnung ohne Beeinflussung der Nutzenden in authentischen Nutzungssituationen (vgl. Kap. 2.1.2, *Textdaten*). In gleicher Perspektive versteht Diekmann (2007: 652) Logdaten als ein Beispiel digitaler Verhaltensspuren und betont die Nutzung des Internet als Instrument zur „Gewinnung nichtreaktiver Daten über Verhaltensweisen und soziale Interaktionen“. Diese Spuren entstehen im Prozess der Auseinandersetzung bzw. im Prozess der Interaktion von Nutzenden mit Online-Umgebungen ohne direkten äußeren Eingriff und werden daher auch als „prozessgenerierte“ bzw. „prozessproduzierte“ Daten be-

zeichnet (vgl. auch Rohwer / Pötter 2002, Baur 2005, Schnell / Hill / Esser 2005, Häder 2006 , Diekmann 2007) oder auch als „elektronische Prozessdaten“ (Bergmann / Meier 2000: 431, vgl. Kap. 2.1.2, *Textdaten*).

Bei der vorgelegten Navigationsanalyse handelt es sich also um eine *explorativ-heuristische Studie;* um die systematische Analyse von *nicht-reaktiven Verhaltensspuren als prozessproduzierten Daten* (Logdaten), die in *authentischen Nutzungssituationen* entstanden sind.[90]

Die empirische Datenbasis der vorgelegten Navigationsanalyse beruht auf den Zugriffen von Nutzenden der hypertextuelle, metadatenbasierte Online-Lernumgebung *Lerndorf* auf den Bereich *Statistik*, genauer auf die Lerneinheiten „Maße der zentralen Tendenz“ (513), „Arithmetisches Mittel“ (514), „Median“ (516) und „Modus“ (517). Analysiert werden Prozesse der *Mikronavigation* als Navigation innerhalb der Lerneinheiten, aus denen die genannten Wissenseinheiten bestehen (vgl. Kap. 3, *Web-Didaktik*).

Im Rahmen der hier durchgeführten Navigationsanalyse werden die Zugriffe im Zeitraum 10/2004 bis 02/2006 berücksichtigt.

Lerneinheit	Anzahl analysierter Wissenseinheiten	Anzahl analysierter Sequenzen
Maße der zentralen Tendenz (513)	1542	475
Arithmetisches Mittel (515)	1638	526
Median (516)	604	212
Modus (517)	896	309
	4680	1522

Tabelle 10: Anzahl der analysierten Wissenseinheiten und Sequenzen im Überblick

Insgesamt beruhen die folgenden Ergebnisse auf der Analyse von insgesamt 1522 Navigationssequenzen, die aus insgesamt 4680 Wissenseinheiten bestehen (vgl. Tab. 10). Die durchschnittliche Sequenzlänge beträgt 3,07 Elemente. Legt man der Berechnung der Sequenzlänge den um die Ein-Element-Sequenzen bereinigten Datensatz zugrunde, beträgt die durchschnittliche Sequenzlänge 4,39 Elemente.

Über diese Zugriffe hinaus stehen weder Daten zur Soziodemographie der Nutzenden noch zur Intention oder zum Erfolg bzw. Misserfolg der Nutzung zur Verfügung.

90 Diese Verhaltensspuren in Form von Logdaten entstehen dabei nicht auf *natürliche* Weise bzw. von selbst. Wesentlich hierbei ist die explizite Definition der Aufzeichnung sowie der Art und Weise der Aufzeichnung der Logdaten (vgl. Kap. 2.1, *Aufzeichnungsverfahren)*. So weisen Schnell / Hill / Esser (2005) darauf hin, dass auch prozessproduzierte Daten einen Entstehungsprozess haben, der sich auf die Güte der Daten auswirken kann.

10.2 Konzeptionell-theoretische Durchführung

In diesem Abschnitt wird die Durchführung der Navigationsanalyse auf konzeptionell-theoretischer Ebene zusammenfassend dargestellt. Dazu wird am konkreten Beispiel der TDA-Syntax die Durchführung der Optimal-Matching Analyse anhand des *seqm*-Befehls und dessen Parametern erläutert (vgl. Abb. 49: 134): Generell wird die Optimal-Matching Analyse auf der Grundlage der Operationen Einfügen („insertion"), Löschen („deletion") und Ersetzen („substitution") durchgeführt.

- Die Indelkosten (*icost*) sowie die Substitutionskosten (*scost*) werden entsprechend der default-Einstellung definiert. Diese Entscheidung wurde in Kapitel 9.1, *Definition der Substitutionskosten im Rahmen der Optimal-Matching Analyse* ausführlich begründet.
- Das zwei- oder mehrfache Auftreten eines identischen Elementes (Wissenseinheit) unmittelbar hintereinander in den Sequenzen (Doppeleintrag) wird bei der Analyse durch ein einmaliges Auftreten ersetzt. Dazu wird innerhalb des *seqm*-Befehls in TDA der Parameter *sm* mit „2" definiert. Aus inhaltlich-technischen Gründen werden diese Doppeleinträge bei der vorliegenden Navigationsanalyse nicht berücksichtigt, da diese aus Perspektive des Nutzers nicht mit der Darstellung neuer Informationen bzw. Wissenseinheiten verbunden ist. Die dargestellte Wissenseinheit ist die gleiche, unabhängig vom Doppel- oder Einfacheintrag in den Logdaten. Darüber hinaus entstehen solche doppelten Einträge lediglich in sehr seltenen, technisch bedingten Fällen.[91]
- Mit dem Parameter *dtda* wird eine Datei erzeugt, die eine Beschreibung des formalen Aufbaus der TDA-Ausgabedatei enthält. Sie dokumentiert die Struktur der Datei, die das Ergebnis der Optimal-Matching Analyse enthält. Es wurde darauf hingewiesen, dass TDA aus Gründen der Darstellbarkeit die errechnete Distanzmatrix in Form einer Tabelle dokumentiert: Jede Zeile dieser Tabelle enthält die Dokumentation eines paarweisen Sequenzvergleichs. Die erste Spalte enthält die Fallnummer der jeweiligen Ausgangssequenz, die zweite Spalte die Fallnummer der jeweiligen Zielsequenz, die dritte Spalte die Anzahl der Elemente der Ausgangssequenz, die vierte Spalte die Anzahl der Elemente der Zielsequenz und die fünfte Spalte die Levenshtein-Distanz der Ausgangs- und der Zielsequenz.

Darüber hinaus ist hervorzuheben, dass bei der Durchführung der vorliegenden Navigationsanalyse der Navigationsprozess als *zeitlicher Verlauf* der Abfolge von Wissenseinheiten analysiert wird. Nicht berücksichtigt wird dabei die *Dauer* des Aufenthalts in der betreffenden Wissenseinheit (Verweildauer).

91 Das Auftreten doppelter Einträge innerhalb der Sequenzen tritt nur in dem empirisch seltenen Fall auf, dass die Auswahl von Verknüpfungen mit einem Doppelklick statt eines einfachen Klicks erfolgt, und gleichzeitig dieser Doppelklick in einer eng begrenzten Zeitspanne erfolgt. Bei Versuchen der Generierung solcher Doppeleinträge konnten keine längeren Ketten erzeugt werden, da der zeitliche Rahmen, in dem der Doppelklick erfolgen muss, im Bereich von Millisekunden liegt. Sind in den Sequenzdaten längere Ketten dieser Doppeleinträge erkennbar, kann davon ausgegangen werden, dass es sich nicht um einen menschlichen sondern um einen maschinellen „Nutzer" (Robot, Spider, u.ä.) handelt. Gleichwohl können diese Doppeleinträge bei der Analyse aggregierter oder sequenzierter Logfiles identifiziert und insofern berücksichtigt werden, dass sie bei der Analyse menschlicher Nutzer ausgeschlossen werden.

Eine *Standardisierung* der Sequenzen und vor allem der Sequenzlänge wird nicht durchgeführt: Alle Zugriffe auf die betreffende Lerneinheit werden dokumentiert und in der Sequenzanalyse berücksichtigt. Das Analysefenster umfasst dabei den Zeitraum den Navigationsverlauf *innerhalb* der definierten Lerneinheit (Mikronavigation). Das Analysefenster schließt auch einmalige Aufrufe *einer* Wissenseinheit ein, die auf Navigationsverläufe der Makroebene verweisen.

Die im Rahmen der Optimal-Matching Analyse erzeugte Datei mit den paarweisen Distanzen der Sequenzen bildet dann den Ausgangspunkt einer Clusteranalyse. Dazu werden die Distanzen mit dem Programm TDA auf Grundlage des Ward-Verfahrens geclustert und die Lösungen mit 10 bzw. 28 Clustern interpretiert. Die ausführliche Diskussion der Wahl des Clusterverfahrens wurde in Kapitel 9.3.5 dargestellt; sowie die Diskussion der Anzahl der Cluster in Kapitel 9.4. Die Clusterlösung als Ergebnis der Clusteranalyse bildet dann den Ausgangspunkt für die formale und inhaltliche Interpretation unterschiedlicher Gruppen von Navigationsweisen.

10.3 Programmtechnische Durchführung

In diesem Absatz werden programmtechnische Aspekte der Durchführung der Navigationsanalyse zusammenfassend dargestellt, die für die Durchführung der Navigationsanalyse eine besondere Herausforderung darstellen.

Dazu wird sowohl das Vorgehen der Datenaufzeichnung und -aufbereitung (Preprocessing) beschrieben, das auf der Ebene der Daten erst die Voraussetzung für die Optimal-Matching Analyse bildet, als auch die im Rahmen der Navigationsanalyse verwendeten Programme dargestellt. Da gegenwärtig keine umfassende programmtechnische Lösung zur Durchführung der Navigationsanalyse existiert, ist der Einsatz unterschiedlicher Software-Programme notwendig.

Auf der Ebene der Datenaufzeichnung und -aufbereitung stellt sich die Durchführung der Navigationsanalyse als Abfolge unterschiedlicher Schritte dar. Die Navigation von Nutzern innerhalb der Lernumgebung wird auf Grundlage der serverseitigen Logdaten mit Hilfe eines Skriptes (Perl) in einer SQL-Datenbank aufgezeichnet. Diese Aufzeichnung wird automatisiert für jeden Zugriff durchgeführt: Für jeden angemeldeten Nutzer und für jede Nutzerin[92] wird der Navigationsprozess in der Lernumgebung aufgezeichnet, d.h. jeder Nutzende kann zu jedem Zeitpunkt genau einem Zustand (z.B. einer konkreten Wissenseinheit) zugeordnet werden. Dabei ist diese Zuordnung eindeutig und die Zustände schließen sich gegenseitig aus.

92 Die Nutzung der Lernumgebung setzt eine Anmeldung voraus. Diese Anmeldung ist kostenlos und erfolgt unmittelbar nach der Wahl eines Benutzernamens und eines Passwortes. Durch dieses Verfahren der Anmeldung werden auf der Ebene der Logdaten typische Schwierigkeiten vermieden, wie z.B. die Identifizierung von Nutzern und die Identifizierung von Nutzungsverläufen (*Session*, vgl. Kap. 4.2, *Web-Mining*).

Fokus dieser Aufzeichnung ist die Dokumentation der Aufrufe von Wissenseinheiten als kleinster Analyseeinheit der Navigationssequenz. Im Unterschied zu den in gängigen Logfiles aufgezeichneten Daten[93] werden darüber hinaus vor allem Metadaten der aufgerufenen Einheiten in einer SQL-Datenbank dokumentiert. Konkret beziehen sich die Metadaten auf folgende Informationen (vgl. Anhang Kap. 17.5):

- Fortlaufende Fallnummer des Eintrags;
- eindeutige Nutzerkennung (kenn_id);
- Datum des Zugriffs (datum);
- Uhrzeit des Zugriffs (zeit);
- Dauer des Zugriffs (dauer_sek);
- Kennung der Wissenseinheit (art_id);
- Kennung der Wissensart (art_id);
- Name der Wissensart (art_txt);
- Medientyp (medi);
- Name des Medientyps (medi_txt);
- Kennung des Kurses, falls Wissenseinheit innerhalb eines Kurses aufgerufen wurden (kurs);
- Name des Kurses, falls Wissenseinheit innerhalb eines Kurses aufgerufen wurden(kurs_txt);
- Übergeordnetes Gebiet der Wissenseinheit (gebiet);
- Name des übergeordneten Gebietes (gebiet_txt);
- Übergeordneter Bereich der Wissenseinheit (bereich);
- Name des übergeordneten Bereiches (bereich_txt);
- Kennung der Lerneinheit (thema);
- Name der Lerneinheit (them_txt);
- (modus);
- (navigat);
- (lexikon).

Aus dieser Datenbank der Zugriffe auf die Lernumgebung werden per Datenbankabfrage die Zugriffe in Form von Sequenzen generiert. Diese Sequenzen enthalten alle Zugriffe von definierten Nutzern und Nutzerinnen auf spezifische, definierte Lerneinheiten (z.B. Lerneinheit 513, *Maße der zentralen Tendenz*). Die Sequenz beginnt mit dem ersten Zugriff auf eine Wissenseinheit, die der Lerneinheit (513) zugeordnet ist. Die Sequenz endet, sobald eine Wissenseinheit aufgerufen wird, die *nicht* der Lerneinheit 513 zugeordnet ist.

Die Dokumentation der Zugriffe in sequenzierter Form hat dabei folgendes Format: jede Zeile enthält genau eine Sequenz; die einzelnen Elemente der Sequenz sind durch ein Leerzeichen getrennt und am Ende jeder Zeile befindet sich ein Zeilenumbruch.

93 Zur Definition von Logfiles, vgl. <http://httpd.apache.org/docs/1.3/logs.html>, (28.08.2006).

Die Datenbankabfrage zur Generierung der Navigationssequenzen innerhalb einer Lerneinheit kann am Beispiel der oben dargestellten Datenbankstruktur folgendermaßen beschrieben werden: In einem ersten Schritt wird für die erste definierte Nutzerin (d.h. für jede definierte kenn_id) abgefragt, ob die Kennung einer bestimmten *Lerneinheit* (thema, z.B. „513“) in den jeweiligen Logdaten vorhanden ist. Wenn dies der Fall ist, wird die Kennung der entsprechenden *Wissenseinheit* (art_id, z.B. „2642“) in ein Ausgabedokument geschrieben. In einem zweiten Schritt wird ermittelt, ob die folgende, von der gleichen Nutzerin aufgerufene Wissenseinheit ebenfalls den Eintrag der entsprechenden *Lerneinheit* (thema) enthält. Ist dies der Fall, wird auch die Kennung dieser Wissenseinheit in das Ausgabedokument geschrieben, und zwar in die gleiche Zeile wie die vorangehende Kennung und durch ein Leerzeichen getrennt. Enthält die nächste aufgerufene Wissenseinheit *nicht* die Kennung der entsprechenden Lerneinheit, wird in das Ausgabedokument kein weiterer Eintrag vorgenommen, sondern in die betreffende Zeile ein Umbruch eingefügt. Die Logdaten werden dann nach dem nächsten Eintrag der Kennung der *Lerneinheit* durchsucht. Ist in den Logdaten kein Eintrag der Kennung der Lerneinheit (mehr) vorhanden, wird zur nächsten definierten Nutzerkennung (kenn_id) übergegangen und der Prozess startet von neuem.

Auf diese Weise werden per Datenbankabfrage aus den Logfiles Sequenzen abgerufen, die den Navigationsverlauf von Nutzern und Nutzerinnen in der Lernumgebung dokumentieren. Im Fall der vorliegenden Navigationsanalyse werden die Navigationssequenzen von Nutzenden *innerhalb* von Lerneinheiten dokumentiert und analysiert. Diese Navigationssequenzen wurden in Kapitel 3.2 als Mikronavigation beschrieben.

Die auf diese Weise mit Hilfe eines Skriptes abgerufenen Datenbankeinträge in Form von Sequenzen werden mit GREP[94] aufbereitet. So werden insbesondere zur Detailanalyse die Sequenzen unterschiedlicher Nutzerinnen (kenn_id) mit GREP nach bestimmten Elementen durchsucht sowie im Rahmen der Datenaufbereitung und Datenfilterung die Einträge der Projektmitarbeiter der Lernumgebung aus dem zu analysierenden Datensatz gefiltert. Das Dokument mit den aus der Datenbank erzeugten Sequenzen wird dann in SPSS als Textdatei importiert. SPSS dient im Rahmen der Navigationsanalyse sowohl zum *Verwalten* der Sequenzdaten, als auch zu deren *Analyse* (Korrelationen, Kreuztabellen).

Die bisherigen Schritte dienten der Datenaufbereitung. Der folgende Schritt besteht aus der Optimal-Matching Analyse als Kern der Navigationsanalyse: Die Optimal-Matching Analyse wird mit der Software TDA („Transition Data Analysis“) durchgeführt. Als Ergebnis der Optimal-Matching Analyse liegt dann eine Datei vor, die für jeden paarweisen Sequenzvergleich die Levenshtein-Distanz enthält.

Auf Grundlage dieser Datei mit der Dokumentation der Levenshtein-Distanzen wird dann in TDA eine Clusteranalyse nach Ward durchgeführt. Das Ergebnis dieser Clusteranalyse ist die Gruppierung der Sequenzen nach *Ähnlichkeit,* d.h. für jede Sequenz wird eine neue Information in Form der Zugehörigkeit zu einem spe-

94 GREP („Global search for a regular expression and print out matched lines“) ist ein Programm, mit dem Dateien nach definierten Zeichenfolgen („regular expressions“) durchsucht und sortiert werden können; vgl. <http://www.gnu.org/software/grep/>, (28.08.2006).

zifischen Cluster erzeugt. Zur grafischen Darstellung und Beschreibung des Agglomerationsprozesses der Clusteranalyse werden mit dem Programm *ClustanGraphics* Dendogramme erzeugt. Die spezifische Clusterzugehörigkeit wird als neue Variable in den SPSS-Datensatz eingefügt. Der Datensatz wird dann entsprechend der Variable *Clusterzugehörigkeit* sortiert, um eine Darstellung der in den unterschiedlichen Clustern enthaltenen Sequenzen zu erhalten.

Eine besondere Herausforderung bei der programmtechnischen Durchführung der Navigationsanalyse stellen die unterschiedlichen Ein- und Ausgabeformate der verwendeten Programme dar: So ist z.B. ein *Umformatieren* der Datenstruktur erforderlich, um das Ergebnis der Optimal-Matching Analyse von TDA zur weiteren Analyse in ClustanGraphics bearbeiten zu können.[95]

Eine weitere Herausforderung besteht in der Bedienung von TDA als kommandozeilenorientiertem Programm ohne grafische Benutzeroberfläche (GUI) und mit eigenwilliger Syntax, vor allem was die Erzeugung und Darstellung von Grafiken betrifft (vgl. Rohwer / Pötter 2005).

10.3.1 Validierung der technischen Erhebung der Sequenzdaten

Da es sich bei der beschriebenen Datenaufzeichnung und Datenaufbereitung (Preprocessing) um ein komplexes Vorgehen handelt, bildet die Validierung dieser Prozesse eine wesentliche Voraussetzung für die weitere Analyse: Arbeitet der Prozess der Datenaufzeichnung und Datenaufbereitung technisch fehlerfrei? Entsprechen die aufgezeichneten und aufbereiteten Sequenzen den konkreten Navigationsverläufen der Nutzenden?

Zur Beantwortung dieser Fragen wurde in der Lernumgebung zunächst auf eine genau definierte Weise navigiert, wobei insbesondere die Abfolge von Wissenseinheiten (und die Abfolge der entsprechenden Lerneinheiten) dokumentiert wurde. Darüber hinaus wurde diese Navigation mit Hilfe einer Screen-Recording Software aufgezeichnet. Anschließend wurden per Datenbankabfrage für die betreffende Nutzerkennung die Navigationssequenzen innerhalb der entsprechenden Lerneinheiten abgerufen.

Die auf Grundlage der Datenbankabfrage erhobenen Sequenzen wurden mit den dokumentierten Navigationsweisen verglichen. Dabei wurden sowohl die Sequenzen als Abfolge von Wissenseinheiten als auch der Wechsel zwischen Lerneinheiten überprüft. Zusätzlich wurde der zeitliche Verlauf des Navigationsprozesses verglichen.

Das abschließende Ergebnis dieser Validierung zeigt, dass die dokumentierten Navigationssequenzen mit denen auf Grundlage der Datenbankabfrage übereinstimmen, was die Abfolge wie auch den zeitlichen Verlauf betrifft: Das Vorgehen zur Datenaufzeichnung und Datenaufbereitung ist damit technisch korrekt und valide.

95 Für die Diskussion zahlreicher Fragen der Formatierung, des Im- und Exportes sowie der generellen Nutzung von TDA geht mein besonderer Dank an meinen Kollegen Heinz-Günther Micheel.

10.3.2 Syntax der Optimal-Matching Analyse (TDA)

In diesem Abschnitt wird die konkrete programmtechnische Umsetzung der Optimal-Matching Analyse anhand der TDA-Befehlssyntax und deren Parameter dargestellt.

Ausgangspunkt der Durchführung der Optimal-Matching Analyse ist ein Dokument als Ergebnis der Datenaufzeichnung und Datenaufbereitung (in diesem Fall im *.txt Format), das die zu analysierenden Sequenzen enthält. Ergebnis der Optimal-Matching Analyse ist ein Dokument, dass für jeden paarweisen Sequenzvergleich die Levenshtein-Distanz enthält.

Zur programmtechnischen Durchführung der Optimal-Matching Analyse mit TDA sind folgende Schritte erforderlich:

1. Konvertierung der Ausgangsdatei (*.txt) in eine SPSS-Datei (*.sav). und Erzeugen einer eindeutigen Fallnummer für jede Sequenz des Datensatzes.
2. Einlesen der SPSS-Datei in TDA (*rspss1, v*gl. Abb. 46).

```
rspss1 (
    noc=...,        maximum number of cases, def. all
    msys=...,       system missing value code, def. -5
    df=...,         write data directly to output file
    dvar=...,       create file with variable descriptions
) = file_name;

wspss1 (
```

Abbildung 46: Syntax des rspss1-Befehls (TDA-Manual 2005: 89).

3. Erstellen der TDA-Datenstruktur und der Variablen (*nvar*, vgl. Abb. 47).

```
nvar (
   vdef,          definition of a variable; this parameter can be used
                  several times to define several variables simultane-
                  ously
   noc=...,       maximum number of cases, def. 1000
   isc=...,       separation character between entries
   dfile=...,     definition of an external data file
   dreclen=...,   fixed length of data file records
   nmrec=...,     number of multiple records, def. 1
   ffmt=...,      format information for data file records
   match=...,     information about matching variables
   isel=...,      case selection while reading data
   vsel=...,      case selection while creating variables
   break=...,     break on condition
   dblock=...,    definition of block mode
   bsel=...,      block mode case selection
   fmt=...,       new print format
   arcdic,        shows value labels for archive variables
   mblnk=...,     new missing value code: blanks, def. -1
   mstar=...,     new missing value code: stars, def. -1
   mpnt=...,      new missing value code: points, def. -1
   mmatch=...,    new missing value code: mismatches, def. -3
   mgen=...,      new missing value code: general, no default
   df=...,        creates an output data file
   keep=...,      keep variables for df option
   drop=...,      drop variables for df option
   bsize=...,     maximum block size, def. 1000
   dtda=...,      TDA description file
   dspss=...,     SPSS description file
);
```

Abbildung 47: Syntax nvar: TDA-Manual (2005; 42).

4. Definition der Sequenz (*seqdef*, vgl. Abb. 48).

```
seqdef (
    sn=...,      number of sequence data structure, def. 1
    m=...,       type of input data, def. 1
    rc=...,      option for recoding states
) = varlist;
```

Abbildung 48: Syntax des seqdef-Befehls (TDA-Manual 2005; 140)

5. Durchführung der Optimal-Matching Analyse als eines paarweisen Vergleichs aller im Datensatz enthaltenen Sequenzen (*seqm,* vgl. Abb. 49).

   Der *seqm*-Befehl stellt den Kern der Optimal-Matching Analyse dar. Insbesondere werden durch die Parameter die Indelkosten (*icost*) und Substitutionskosten (*scost)* definiert. Mit dem Parameter *df* wird eine Datei erzeugt („test output file“), in der die verwendeten Indelkosten und Substitutionskosten dokumentiert werden. Der Parameter *tst* stellt eine Spezifizierung des „test output file“ dar:

durch Wahl der Option „3“ wird für jeden paarweisen Sequenzvergleich der Optimal-Matching Analyse die jeweilige Matrix zur Ermittlung der Levenshtein-Distanz eingefügt (Eine ausführliche Dokumentation des „test output file“ sowie der Spezifizierung durch das *tst*-Parameter befindet sich in Kapitel 7.1).

```
seqm (
    m=...,        selection of method, def. 1
    sn=...,       selection of sequence data structure(s), def. 1
    icost=...,    indel cost specification
    scost=...,    substitution cost specification
    rr=1,         use common sequence length
    sm=...,       option for preprocessing sequences
    r=...,        random selection of sequences
    s=...,        print sequence of distances, or LCS
    cn=...,       compare with specified sequences
    max=...,      alignment restriction
    tfmt=...,     print format for distances, def. 5.2
    v=...,        add variables ... to output file
    dtda=...,     create TDA description file
    df=...,       create test output file
    tst=...,      additional test options
    fmt=...,      print format for test output file
) = fname;
```

Abbildung 49: Syntax des seqm-Befehls (TDA-Manual 2005; 480).

11 Ergebnisse der Navigationsanalyse

In diesem Kapitel werden die Ergebnisse der Navigationsanalyse dargestellt. Den Ausgangspunkt bilden die spezifischen Cluster als Ergebnis der Clusteranalyse, in denen die empirischen Sequenzen des Datensatzes nach dem Kriterium der Levenshtein-Distanz gruppiert vorliegen.

Dazu wird in einem ersten Schritt das Ergebnis der Clusteranalyse beschrieben: Der Agglomerationsprozess der Clusteranalyse wird anhand eines Dendogramms sowie die Zunahme des Fusionskoeffizienten im Verlauf der Clusteranalyse grafisch anhand eines Häufigkeitspolygons verdeutlicht. Die konkreten Häufigkeitsverteilungen der Cluster für die Lösungen mit 10 und 28 Clustern werden tabellarisch dargestellt.

Daran anschließend werden die spezifischen Clusterlösungen für jede der vier analysierten Lerneinheiten auf formaler Ebene beschrieben: Welche typischen Abfolgen von Elementen sind in den spezifischen Clustern enthalten? Welche formalen Merkmale kennzeichnen diese typischen Abfolgen?

In einem zweiten Schritt folgt aufbauend auf der formalen Beschreibung eine zusammenfassende *Interpretation* der Clusterlösungen aus *inhaltlicher Perspektive:* Welche Aussagen können über die empirischen Navigationsverläufe getroffen werden? Welche Muster, Regelmäßigkeiten und Strukturen kommen in den geclusterten Navigationsverläufen zum Ausdruck? Welche Strategien der Navigation sind für den Bereich der Mikronavigation identifizierbar?

Abschließend werden die Ergebnisse der Interpretation der Navigationssequenzen zusammengefasst und es wird auf *Weiterentwicklungen* und *Variationsmöglichkeiten* der Navigationsanalyse hingewiesen.

11.1 Formale Darstellung der Navigationssequenzen

In diesem Abschnitt werden die spezifischen Cluster als Ergebnisse der Clusteranalyse für die vier analysierten Lerneinheiten auf *formaler Ebene* beschrieben. Dazu werden die Häufigkeitsverteilungen der Clusterlösung mit 10 und 28 Clustern tabellarisch dargestellt. Der Agglomerationsprozess der Clusteranalyse wird mit Hilfe eines Dendogramms, sowie die Zunahme des Fusionskoeffizienten anhand eines Häufigkeitspolygons verdeutlicht.

Bei der formalen Beschreibung der einzelnen Cluster wird die *Clusterzugehörigkeit* in Bezug auf die Clusterlösung mit 10 und 28 Clustern angegeben. So steht die Bezeichnung „4/10; 14/28“ für das Cluster Nummer 4 bezogen auf die Clusterlösung mit 10 Clustern („Cluster 4 von 10“); gleichzeitig stellt dieses Cluster

das Cluster Nummer 14 bezogen auf die Clusterlösung mit 28 Clustern da („Cluster 14 von 28“). Anhand dieser Doppelbezeichnung ist also nachvollziehbar, wie die Clusterlösungen mit 28 Clustern im weiteren Verlauf der Clusteranalyse fusioniert werden. Gleichzeitig entspricht die Analyse der Clusterlösung mit 28 Clustern einem hohen Grad an Differenzierung und damit dem explorativ-heuristischen Ansatz der Navigationsanalyse.

Darüber hinaus wird für jedes Cluster die Anzahl der Sequenzen angegeben, die in diesem Cluster enthalten sind, sowie der *prozentuale Anteil* an der Anzahl aller Sequenzen in der jeweiligen Lerneinheit. Anhand der eindeutigen *Kennung* der Wissensart (kenn_id), anhand des Metadatums der *Wissensart* (art_id) sowie der *Anzahl der Elemente* dieser typischen Abfolge wird die für jedes Cluster *typische Abfolge* von Wissenseinheiten dargestellt.[96]

Abschließend wird jedes Cluster hinsichtlich formaler *Merkmale* beschrieben: So wird z.B. darauf hingewiesen, ob es sich hinsichtlich der Abfolge um ein homogenes oder heterogenes Cluster handelt; welche Wissenseinheit den Ausgangspunkt der Abfolge darstellt u.ä. Darüber hinaus wird darauf hingewiesen, ob die typische Abfolge eine Navigation „von links nach rechts“ dargestellt. Dieses Merkmal bezieht sich auf die grafische Anordnung (Screendesign) der Wissenseinheiten innerhalb der Lerneinheit (vgl. Abb. 54: 140) und bedeutet, dass die Abfolge der Elemente der Sequenz der grafischen Anordnung der Wissenseinheiten entspricht, diese also in Bezug auf das Layout der Lernumgebung in der Reihenfolge „von links nach rechts“ ausgewählt wurden. Im Gegensatz zu dieser Form der Navigation wird ein Abweichen von dieser Reihenfolge als „direkter Zugriff“ gekennzeichnet. Aus Perspektive der Navigation „von links nach rechts“ werden dabei einzelne oder mehrere Wissenseinheit *ausgelassen* bzw. *übersprungen*. Hinsichtlich der grafischen Darstellung der Wissenseinheiten in der Lernumgebung (Screendesign) kann die Navigation von „von links nach rechts“ auch als *lineare* Navigation und der „direkte Zugriff“ als *nicht-lineare* Navigation bezeichnet werden.

Von der formalen Beschreibung der Cluster ausgeklammert werden Cluster, die eine sehr geringe Anzahl von Sequenzen bzw. lediglich eine Sequenz enthalten. In formaler Hinsicht unterscheiden sich diese Sequenzen von allen anderen Sequenzen des Datensatzes so stark, dass sie im Agglomerationsprozess bis zur Clusterlösung mit 28 Sequenzen mit keinem anderen Cluster fusioniert wurden. Im analysierten Datensatz sind auf der 28 Cluster Ebene insgesamt 15 Cluster mit lediglich einer Sequenz enthalten (*Ein-Sequenz-Cluster*). Die durchschnittliche Länge der Sequenz beträgt dabei 14 Wissenseinheiten; wobei die konkrete Anzahl der Elemente zwischen 7 und 27 Elementen liegt. In der Tendenz handelt es sich bei diesen Sequenzen also um lange bis sehr lange Navigationssequenzen.

Von der weiteren Analyse ausgeklammert werden diese Sequenzen (*Ein-Sequenz-Cluster*), da der Fokus der vorliegenden Navigationsanalyse auf Mustern, Regelmäßigkeiten und Strukturen der Navigationssequenzen

96 Die *typische Abfolge* der Wissenseinheiten des Clusters wird dabei interpretativ bestimmt. Eine mathematische Bestimmung beispielsweise durch eine Clusterzentrenanalyse ist nicht möglich, da über die Distanzmatrix hinaus keine weiteren Variablen zur Berechnung des Clusterzentrums zur Verfügung stehen.

liegt und anhand einzelner Sequenzen solche clusterübergreifenden Muster nicht erkannt werden können, da diese erst in der „Gesamtschau“ (Erzberger 2001: 136, vgl. Kap. 6.7) einer hinreichend großen Anzahl von Sequenzen identifizierbar werden. Grundsätzlich ist jedoch im Rahmen der Navigationsanalyse auch eine Fokussierung auf diese *Ein-Sequenz-Cluster* möglich. Die weitergehende Analyse von Clustern mit sehr geringer Fallzahl erhält besonders im Rahmen einer *Triangulation* analytisches Potenzial; z.B. im Zusammenhang mit retrospektiven fokussierten Interviews oder aber auch im Zusammenhang mit der Methode *Lauten Denkens* (vgl. Kap. 2, *Methodologische Grundlagen der Navigationsanalyse*).

11.1.1 Lerneinheit 513: Maße der zentralen Tendenz

Das folgende Dendogramm (vgl. Abb. 50) beschreibt den Agglomerationsprozess der Clusteranalyse für die Sequenzen der Lerneinheit „Maße der zentralen Tendenz“ (513). Ausgangspunkt ist die Clusterlösung mit 28 Clustern, die in der linken Spalte dargestellt werden. Anhand des Dendogramms wird deutlich, welche Cluster in den folgenden Schritten fusioniert werden. Darüber hinaus verdeutlicht die Länge der Linien auf der x-Achse die Zunahme des Fusionskoeffizienten bei fortschreitender Fusionierung.

Die Zunahme des Fusionskoeffizienten bei der Fusionierung von Clustern wird anhand folgender Grafik deutlich (vgl. Abb. 51). Im Verlauf der Clusteranalyse (Ward) werden zu Beginn solche Cluster fusioniert, die zu einer geringen Zunahme der Fehlerquadratsumme führen. Im Verlauf der Clusteranalyse steigt dieser Fusionskoeffizient immer weiter an, da Cluster fusioniert werden, die sich immer weniger ähnlich sind.

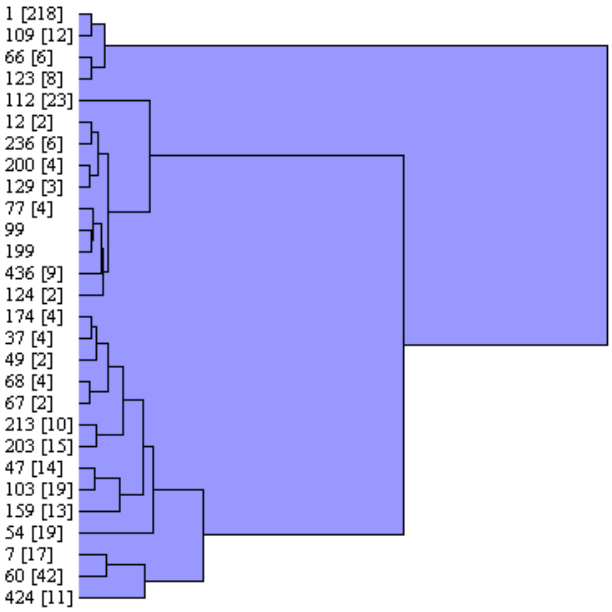

Abbildung 50: Clusterlösung der Lerneinheit „Maße der zentralen Tendenz“ (513)

Für die Daten der Lerneinheit „Maße der zentralen Tendenz“ beträgt der t-Wert für die Clusterlösung mit 28 Clustern 8,0 und für die Clusterlösung mit 10 Clustern 29,5. Statistisch signifikant auf einem 5% Niveau sind alle Clusterlösungen mit weniger als 56 Clustern.

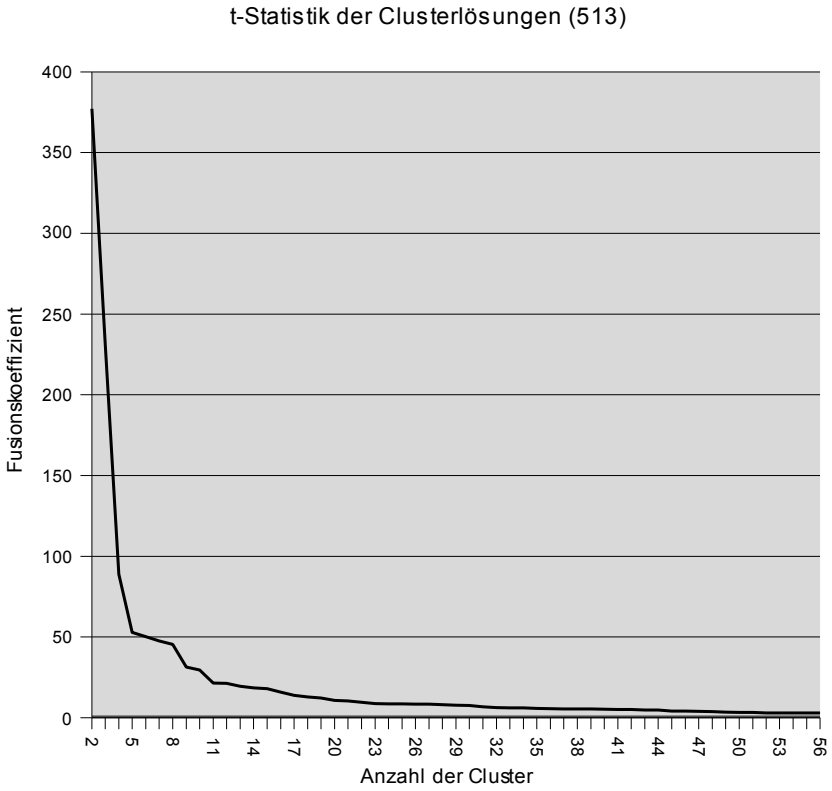

Abbildung 51: t-Statistik der Clusterlösungen, „Maße der zentralen Tendenz“ (513).

Für einen Überblick über das Ergebnis der Clusteranalyse wird im Folgenden die Häufigkeitsverteilung für die Clusterlösung mit 10 Clustern (vgl. Abb. 52) und mit 28 Clustern (vgl. Abb. 53) tabellarisch dargestellt. Diese tabellarische Darstellung dient als Ausgangspunkt für die folgende formale Beschreibung der Sequenzen der Cluster.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	244	51,4	51,4	51,4
	2	23	4,8	4,8	56,2
	3	32	6,7	6,7	62,9
	4	16	3,4	3,4	66,3
	5	25	5,3	5,3	71,6
	6	33	6,9	6,9	78,5
	7	13	2,7	2,7	81,3
	8	19	4,0	4,0	85,3
	9	59	12,4	12,4	97,7
	10	11	2,3	2,3	100,0
	Gesamt	475	100,0	100,0	

Abbildung 52: Häufigkeitsverteilung der Ward-Clusterlösung mit 10 Clustern (Lerneinheit 513, „Maße der zentralen Tendenz“).

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	218	45,9	45,9	45,9
	2	12	2,5	2,5	48,4
	3	6	1,3	1,3	49,7
	4	8	1,7	1,7	51,4
	5	23	4,8	4,8	56,2
	6	2	,4	,4	56,6
	7	6	1,3	1,3	57,9
	8	4	,8	,8	58,7
	9	3	,6	,6	59,4
	10	4	,8	,8	60,2
	11	1	,2	,2	60,4
	12	1	,2	,2	60,6
	13	9	1,9	1,9	62,5
	14	2	,4	,4	62,9
	15	4	,8	,8	63,8
	16	4	,8	,8	64,6
	17	2	,4	,4	65,1
	18	4	,8	,8	65,9
	19	2	,4	,4	66,3
	20	10	2,1	2,1	68,4
	21	15	3,2	3,2	71,6
	22	14	2,9	2,9	74,5
	23	19	4,0	4,0	78,5
	24	13	2,7	2,7	81,3
	25	19	4,0	4,0	85,3
	26	17	3,6	3,6	88,8
	27	42	8,8	8,8	97,7
	28	11	2,3	2,3	100,0
	Gesamt	475	100,0	100,0	

Abbildung 53: Häufigkeitsverteilung der Ward-Clusterlösung mit 28 Clustern (Lerneinheit 513, „Maße der zentralen Tendenz“).

Die Anzahl der analysierten Sequenzen in dieser Lerneinheit (513) beträgt insgesamt 475, mit insgesamt 1542 Wissenseinheiten.

Abbildung 54: Wissenseinheit „Maße der zentralen Tendenz“ (513), Kennung – Wissensart.

Im Folgenden werden die Clusterlösungen tabellarisch dargestellt und formal beschrieben:

Cluster 1/10; 1/28

Anzahl der Sequenzen	218
Prozentualer Anteil	45,9%
typische Abfolge:	
- Kennung (art_id)	2642
- Wissensart (art)	Orientierung / Text
Anzahl der Elemente	1

Merkmale Sequenz	• enthält keine weiteren Wissenseinheiten • sehr homogenes Cluster

Cluster 1/10; 2/28

Anzahl der Sequenzen	12
Prozentualer Anteil	2,5 %
typische Abfolge	
Kennung (art_id)	2642 – 2900
Wissensart (art)	Orient/T - Handlung
Kennung (art_id)	2642 - 2999
Wissensart (art)	Orient/T – Aufgabe
Kennung (art_id)	2642 – 3065 – 2642
Wissensart (art)	Orient/T - Literatur - Orient/T
Anzahl der Elemente	2 bzw. 3
Merkmale	• heterogenes Cluster • gemeinsamer Startpunkt: Orientierungswissen / Text • kein Beispielwissen, Erklärungswissen oder Wechsel des Medientyps

Cluster 1/10; 3/28

Anzahl der Sequenzen	6
Prozentualer Anteil	1,3%
typische Abfolge	
Kennung (art_id)	3365
Wissensart (art)	Diskussion
Anzahl der Elemente	1
Merkmale	• enthält keine weiteren Wissenseinheiten der Lerneinheit • Aufruf der Wissenseinheit im Rahmen der Kursnavigation

Cluster 1/10; 4/28

Anzahl der Sequenzen	8
Prozentualer Anteil	1,7%
typische Abfolge	

Kennung (art_id)	2642 – 2998 – (2642; 3365)
Wissensart (art)	Orient./T - Entdeckende Aufgabe – (Orient./T; Orient./A)
Anzahl der Elemente	3
Merkmale	• kein Handlungswissen, Erklärungswissen, Beispielwissen, Quellenwissen • kein Wechsel des Medientyps beim Orientierungswissen • homogenes Cluster • direkter Zugriff auf Entdeckende Aufgabe (und zurück zum Ausgangspunkt)

Cluster 2/10; 5/28

Anzahl der Sequenzen	23
Prozentualer Anteil	4,8%
typische Abfolge	
Kennung (art_id)	2642 – 2900 – 2896 – 3065 – 2999 – 2998 – 3365 – 2642
Wissensart (art)	Orient./T – Handlung – Erklärung – Quellen - Aufgabe – Entd. Aufgabe – Diskussion - Orient./T
Anzahl der Elemente	8
Merkmale	• es werden alle Wissenseinheiten der Lerneinheit aufgerufen, in der Reihenfolge der Anordnung („von links nach rechts“-Navigation). • Die Sequenz endet mit dem Sprung zurück zum Starpunkt (Orientierungswissen) • kein Wechsel des Medientyps beim Orientierungswissen

Cluster 3/10; 6/28

Anzahl der Sequenzen	2
Prozentualer Anteil	0,4%

Cluster 3/10; 7/28

Anzahl der Sequenzen	6
Prozentualer Anteil	1,3%
typische Abfolge	
Kennung (art_id)	3597 – 2900 – 2896 – 2999 – 2898 - 3365
Wissensart (art)	Orient./A - Handlung – Erklärung – Aufgabe – Entd. Aufgabe - Diskussion
Anzahl der Elemente	5
Merkmale	• Navigation „von links nach rechts“ • kein Wechsel des Medientyps beim Orientierungswissen • kein Aufruf Quellenwissen • kein Sprung zurück zum Orientierungswissen (Startpunkt) am Ende der Sequenz, (vgl. Cluster 2/10;

	5/28).

Cluster 3/10; 8/28

Anzahl der Sequenzen	4
Prozentualer Anteil	0,8%
typische Abfolge	
Kennung (art_id)	3597 – 2900 – 2896 – ... - 3597
Wissensart (art)	Orient./A - Handlung – Erklärung – ...
Anzahl der Elemente	12
Merkmale	• die ersten 3 Wissenseinheiten folgen einer Navigation „von links nach rechts“, dann jedoch sehr heterogener weiterer Verlauf: mehrmaliger Wechsel zwischen den Wissenseinheiten der Lerneinheit. • Sprung zurück zum Orientierungswissen (Startpunkt) am Ende der Sequenz, (vgl. Cluster 2/10; 5/28) • kein Wechsel des Medientyps beim Orientierungswissen

Cluster 3/10; 9/28

Anzahl der Sequenzen	3
Prozentualer Anteil	0,6%
typische Abfolge	
Kennung (art_id)	3597 – 2900 – 2896 – 3065 - 2999 – 2998 – 3365 - ...
Wissensart (art)	Orient./A - Handlung – Erklärung – Quellen – Aufgabe – Entd. Aufgabe – Diskussion - ...
Anzahl der Elemente	11
Merkmale	• die Abfolge der ersten 7 Wissenseinheiten entspricht einer Navigation „von links nach rechts“, dann Auswahl einzelner Wissenseinheiten • kein Sprung zurück zum Orientierungswissen (Startpunkt) am Ende der Sequenz, (vgl. Cluster 2/10; 5/28) • kein Wechsel des Medientyps beim Orientierungswissen

Cluster 3/10; 10/28

Anzahl der Sequenzen	4
Prozentualer Anteil	0,8%
typische Abfolge	
Kennung (art_id)	3597 – 2900 – 2896 – 3065 - ...
Wissensart (art)	Orient./A - Handlung – Erklärung – Quellen – ...
Anzahl der Elemente	11
Merkmale	• die Abfolge der ersten 4 Wissenseinheiten entspricht einer Navigation „von links nach rechts“, dann

	Auswahl unterschiedlicher Wissenseinheiten • kein Sprung zurück zum Orientierungswissen (Startpunkt) am Ende der Sequenz, (vgl. Cluster 2/10; 5/28) • kein Wechsel des Medientyps beim Orientierungswissen

Cluster 3/10;11/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,2%
Anzahl der Elemente	14

Cluster 3/10; 12/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,2%
Anzahl der Elemente	15

Cluster 3/10; 13/28

Anzahl der Sequenzen	9
Prozentualer Anteil	1,9%
typische Abfolge	
Kennung (art_id)	3597 – 2900 – 2896 – 2999 – (2900; 2896)
Wissensart (art)	Orient./A - Handlung – Erklärung – Aufgabe – (Handlung; Erklärung)
Anzahl der Elemente	6
Merkmale	• die Abfolge der ersten 3 Wissenseinheiten entspricht einer Navigation „von links nach rechts“, dann Auswahl Aufgabe, danach Auswahl unterschiedlicher Wissenseinheiten mit Schwerpunkt auf Erklärungswissen und Handlungswissen. • kein Quellenwissen (wird bei Navigation „von links nach rechts“ *übersprungen*) • kein Sprung zurück zum Orientierungswissen (Startpunkt) am Ende der Sequenz, (vgl. Cluster 2/10; 5/28) • kein Wechsel des Medientyps beim Orientierungswissen • heterogenes Cluster

Cluster 3/10; 14/28

Anzahl der Sequenzen	2
Prozentualer Anteil	0,4%

Cluster 3/10; 15/28

Anzahl der Sequenzen	4
Prozentualer Anteil	0,8%
typische Abfolge	
Kennung (art_id)	/
Wissensart (art)	/
Anzahl der Elemente	8
Merkmale	• sehr heterogenes Cluster • kein gemeinsamer Startpunkt: Navigation zwischen Aufgaben, Erklärungs- und Handlungswissen • kein Quellenwissen, keine Diskussion • keine Navigation „von links nach rechts“

Cluster 4/10; 16/28

Anzahl der Sequenzen	4
Prozentualer Anteil	0,8%
typische Abfolge	
Kennung (art_id)	2642 – 3597 – 2642 – 3597 - 2900
Wissensart (art)	Orient./T – Orient./A – Orient./T – Orient./A – Handlung
Anzahl der Elemente	5
Merkmale	• Doppelter Wechsel des Medientyps beim Orientierungswissen zwischen Text und Animation, dann Handlung. • kein Quellenwissen und Diskussion

Cluster 4/10; 17/28

Anzahl der Sequenzen	2
Prozentualer Anteil	0,4

Cluster 4/10; 18/28

Anzahl der Sequenzen	4
Prozentualer Anteil	0,8%
typische Abfolge	
Kennung (art_id)	3365 – 2900 – 2896 – 3597 - 2642
Wissensart (art)	Diskussion – Handlung – Erklärung – Orient./A – Orient./T
Anzahl der Elemente	5

Merkmale	• homogenes Cluster • Sequenzen stammen aus Navigation innerhalb des Kurses (Kursnavigation); • Navigation entspricht dem „Zurücklaufen“ im Kurs an den Anfang • kein Quellenwissen und Aufgabe

Cluster 4/10; 19/28

Anzahl der Sequenzen	2
Prozentualer Anteil	0,4%

Cluster 5/10; 20/28

Anzahl der Sequenzen	10
Prozentualer Anteil	2,1%
typische Abfolge	
Kennung (art_id)	3597 – 2900 – 3597 - ...
Wissensart (art)	Orient./A – Handlung – Orient./A - ...
Anzahl der Elemente	6
Merkmale	• heterogenes Cluster • gemeinsamer Sequenzbeginn mit Orientierung/A, Handlung und Orientierung/A, danach Auswahl weiterer Wissenseinheiten, jedoch keine Diskussion (3365) und kein Quellenwissen (3065)

Cluster 5/10; 21/28

Anzahl der Sequenzen	15
Prozentualer Anteil	3,2
typische Abfolge	
Kennung (art_id)	3597 – 2900
Wissensart (art)	Orient./A - Handlung
Anzahl der Elemente	2
Merkmale	• homogenes Cluster • im Vergleich zu Cluster 20/28 kürzer, aus lediglich 2 Wissenseinheiten bestehend

Cluster 6/10; 22/28

Anzahl der Sequenzen	14
Prozentualer Anteil	2,9%
typische Abfolge	
Kennung (art_id)	2642 – 3597

Wissensart (art)	Orient./T - Orient./A
Anzahl der Elemente	2
Merkmale	• Umschalten des Medientyps beim Orientierungswissen von Text auf Animation • keine weiteren Wissenseinheiten

Cluster 6/10, 23/28

Anzahl der Sequenzen	19
Prozentualer Anteil	4,0%
typische Abfolge	
Kennung (art_id)	2642 - 3597 – (2900 / 2896) - 2642
Wissensart (art)	Orient./T – Orient./A – (Handlung, Erklärung) - Orient./T
Anzahl der Elemente	4
Merkmale	• heterogenes Cluster • Umschalten des Medientyps von Orientierung / Text auf Orientierung / Animation, gefolgt von weiterer Wissenseinheiten: Handlung, Erklärung. Endpunkt der Navigation innerhalb der Lernumgebung ist dann wieder das Orientierungswissen (2642).

Cluster 7/10; 24/28

Anzahl der Sequenzen	13
Prozentualer Anteil	2,7
typische Abfolge	
Kennung (art_id)	2642 – 2900 – 2896
Wissensart (art)	Orient./T – Handlung - Erklärung
Anzahl der Elemente	3
Merkmale	• Navigation „von links nach rechts“ • keine weiteren Wissenseinheiten (Quellen, Entd. Aufgabe, Diskussion)

Cluster 8/10; 25/28

Anzahl der Sequenzen	19
Prozentualer Anteil	4,0%
typische Abfolge	
Kennung (art_id)	2642 – 3597 – 2886 – 2900 – 3365
Wissensart (art)	Orient./T – Orient./A – Erklärung – Handlung - Diskussion
Anzahl der Elemente	5

Merkmale	• homogenes Cluster • Abfolge der Wissenseinheiten entspricht der Abfolge im Kurs „Statistik – Maße der zentralen Tendenz“.

Cluster 9/10; 26/28

Anzahl der Sequenzen	17
Prozentualer Anteil	3,6%
typische Abfolge	
Kennung (art_id)	3597 – 2896
Wissensart (art)	Orient./A - Erklärung
Anzahl der Elemente	2
Merkmale	• direkter Zugriff von Orientierungswissen auf Erklärungswissen, • keine weiteren Wissenseinheiten

Cluster 9/10; 27/28

Anzahl der Sequenzen	42
Prozentualer Anteil	8,8%
typische Abfolge	
Kennung (art_id)	3597
Wissensart (art)	Orient./A
Anzahl der Elemente	1
Merkmale	• in diesem Cluster befinden sich 7 Sequenzen mit der Abfolge Orientierung (5297) – Entdeckende Aufgabe (2998)

Cluster 10/10; 28/28

Anzahl der Sequenzen	11
Prozentualer Anteil	2,3%
typische Abfolge	
Kennung (art_id)	2999 – 2998 - (3065)
Wissensart (art)	Aufgabe – Entd. Aufgabe - (Diskussion)
Anzahl der Elemente	3
Merkmale	• Sequenzbeginn 2999 – 2998 entspricht dem zweitem Teil des Kurses „Statistik – Maße der zentralen Tendenz“, in dem diese Lerneinheit (Maße der zentralen Tendenz) behandelt wird. • kein Handlungswissen und kein Erklärungswissen

11.1.2 Lerneinheit 515: Arithmetisches Mittel

Das folgende Dendogramm (vgl. Abb. 55) beschreibt den Agglomerationsprozess der Clusteranalyse für die Sequenzen der Lerneinheit „Arithmetisches Mittel“ (515). Ausgangspunkt ist die Clusterlösung mit 28 Clustern, die in der linken Spalte dargestellt werden. Anhand des Dendogramms wird deutlich, welche Cluster in den folgenden Schritten fusioniert werden. Darüber hinaus verdeutlichen die Kantenlängen auf der x-Achse die Zunahme des Fusionskoeffizienten bei fortschreitender Fusionierung der Cluster.

Die Zunahme des Fusionskoeffizienten bei der Fusionierung von Clustern wird anhand folgender Grafik deutlich (vgl. Abb. 56): Für die Daten der Lerneinheit „Arithmetisches Mittel“ beträgt der t-Wert für die Clusterlösung mit 28 Clustern 10,33 und für die Clusterlösung mit 10 Clustern 40,79. Statistisch signifikant auf einem 5% Niveau sind alle Clusterlösungen mit weniger als 53 Clustern.

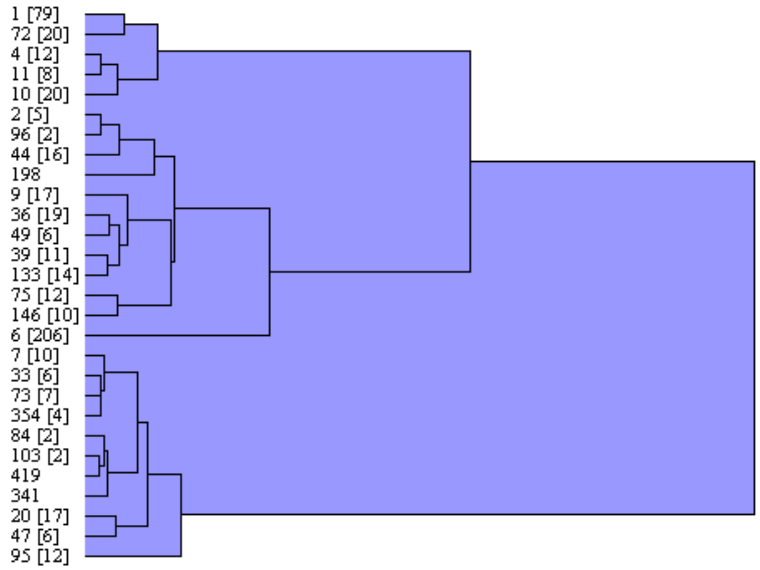

Abbildung 55: Clusterlösung der Lerneinheit „Arithmetisches Mittel“ (515)

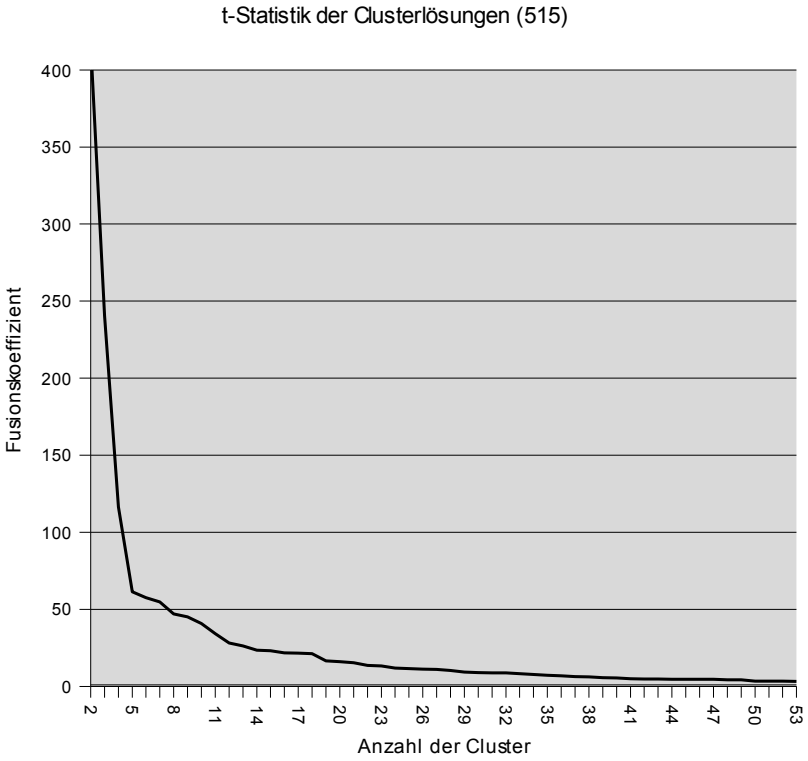

Abbildung 56: t-Statistik der Clusterlösung, „Arithmetisches Mittel" (515).

Für einen Überblick über das Ergebnis der Clusteranalyse wird die Häufigkeitsverteilung für die Clusterlösung mit 10 Clustern (vgl. Abb. 57) und mit 28 Clustern (vgl. Abb. 58) tabellarisch dargestellt. Diese tabellarische Darstellung dient als Ausgangspunkt für die folgende formale Beschreibung der Sequenzen der Cluster.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	99	18,8	18,8	18,8
	2	40	7,6	7,6	26,4
	3	23	4,4	4,4	30,8
	4	1	,2	,2	31,0
	5	67	12,7	12,7	43,7
	6	22	4,2	4,2	47,9
	7	206	39,2	39,2	87,1
	8	33	6,3	6,3	93,3
	9	23	4,4	4,4	97,7
	10	12	2,3	2,3	100,0
	Gesamt	526	100,0	100,0	

Abbildung 57: Häufigkeitsverteilung der Ward-Clusterlösung mit 10 Clustern (Lerneinheit 515, „Arithmetisches Mittel").

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	79	15,0	15,0	15,0
	2	20	3,8	3,8	18,8
	3	12	2,3	2,3	21,1
	4	8	1,5	1,5	22,6
	5	20	3,8	3,8	26,4
	6	5	1,0	1,0	27,4
	7	2	,4	,4	27,8
	8	16	3,0	3,0	30,8
	9	1	,2	,2	31,0
	10	17	3,2	3,2	34,2
	11	19	3,6	3,6	37,8
	12	6	1,1	1,1	39,0
	13	11	2,1	2,1	41,1
	14	14	2,7	2,7	43,7
	15	12	2,3	2,3	46,0
	16	10	1,9	1,9	47,9
	17	206	39,2	39,2	87,1
	18	10	1,9	1,9	89,0
	19	6	1,1	1,1	90,1
	20	7	1,3	1,3	91,4
	21	4	,8	,8	92,2
	22	2	,4	,4	92,6
	23	2	,4	,4	93,0
	24	1	,2	,2	93,2
	25	1	,2	,2	93,3
	26	17	3,2	3,2	96,6
	27	6	1,1	1,1	97,7
	28	12	2,3	2,3	100,0
	Gesamt	526	100,0	100,0	

Abbildung 58: Häufigkeitsverteilung der Ward-Clusterlösung mit 28 Clustern (Lerneinheit 515, „Arithmetisches Mittel“).

Die Anzahl der analysierten analysierten Sequenzen in dieser Lerneinheit (515) beträgt insgesamt 526; die Anzahl der darin enthaltenen Wissenseinheiten 1638.

Abbildung 59: Wissenseinheit „Arithmetisches Mittel“ (515), Kennung – Wissensart.

Im Folgenden werden die Clusterlösungen tabellarisch dargestellt und formal beschrieben:

Cluster 1/10; 1/28

Anzahl der Sequenzen	79
Prozentualer Anteil	15%
typische Abfolge	
Kennung (art_id)	3185
Wissensart (art)	Orient./A
Anzahl der Elemente	1

Merkmale	• enthält keine weiteren Wissenseinheiten der Lerneinheit • in diesem Cluster befinden sich darüber hinaus 3 Sequenzen mit der Abfolge Orientierung (3185) – Quellenwissen (2861); sowie 6 Sequenzen mit der Abfolge Orientierung (3185) – Erklärung (2862).

Cluster 1/10; 2/28

Anzahl der Sequenzen	20
Prozentualer Anteil	3,8%
typische Abfolge	
Kennung (art_id)	3185 - 2885
Wissensart (art)	Orient./A - Handlung
Anzahl der Elemente	2
Merkmale	• Navigation „von links nach rechts“ • kein Wechsel des Medientyps beim Orientierungswissen • enthält keine weiteren Wissenseinheiten der Lerneinheit

Cluster 2/10; 3/28

Anzahl der Sequenzen	12
Prozentualer Anteil	2,3%
typische Abfolge	
Kennung (art_id)	3185 – 2851
Wissensart (art)	Orient./A - Orient./T
Anzahl der Elemente	2
Merkmale	• Wechsel des Medientyps von Animation zu Text, • enthält keine weiteren Wissenseinheiten der Lerneinheit

Cluster 2/10; 4/28

Anzahl der Sequenzen	8
Prozentualer Anteil	1,3%
typische Abfolge	
Kennung (art_id)	2851 – 3185 – 2851 – 3185
Wissensart (art)	Orient./T – Orient./A – Orient./T - Orient./A
Anzahl der Elemente	4
Merkmale	• zweifacher Wechsel des Medientyps von Text zu Animation, • enthält keine weiteren Wissenseinheiten der Lerneinheit

Cluster 2/10; 5/28

Anzahl der Sequenzen	20
Prozentualer Anteil	3,8%
typische Abfolge	
Kennung (art_id)	2851 – 3185
Wissensart (art)	Orient./T - Orient./A
Anzahl der Elemente	2
Merkmale	• einfacher Wechsel des Medientyps von Text zu Animation, • enthält keine weiteren Wissenseinheiten der Lerneinheit • homogenes Cluster

Cluster 3/10; 6/28

Anzahl der Sequenzen	5
Prozentualer Anteil	1,0%
typische Abfolge	
Kennung (art_id)	3185 – 2888 - 2889 – 2863 - 2862
Wissensart (art)	Orient./A – Multiple Choice - True/False – Erklärung - Beispiel
Anzahl der Elemente	5
Merkmale	• Startpunkt Orientierungswissen / Animation, dann direkte Navigation zu Multiple Choice und True/False, danach Beispiel und Erklärung • kein Quellenwissen, kein Handlungswissen

Cluster 3/10; 7/28

Anzahl der Sequenzen	2
Prozentualer Anteil	0,4%

Cluster 3/10; 8/28

Anzahl der Sequenzen	16
Prozentualer Anteil	3,0%
typische Abfolge	
Kennung (art_id)	2851 – 2889 - 2888
Wissensart (art)	Orient./T – True/False - Multiple Choice
Anzahl der Elemente	3
Merkmale	• nach Orientierungswissen / Text direkter Zugriff auf Aufgabe: True/False und Multiple Choice • keine weiteren Wissenseinheiten der Lerneinheit

Cluster 4/10; 9/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,2%
Anzahl der Elemente	27
Merkmale	• sowohl in der Clusterlösung mit 10 wie 28 Clustern bildet diese Sequenz ein einzelnes Cluster, d.h. die Distanz zu allen anderen Sequenzen ist besonders groß. Diese Sequenz ist mit insgesamt 28 Elementen die längste Sequenz dieses Datensatzes.

Cluster 5/10; 10/28

Anzahl der Sequenzen	17
Prozentualer Anteil	3,2%
typische Abfolge	
Kennung (art_id)	2851 - 2863
Wissensart (art)	Orient./T - Beispiel
Anzahl der Elemente	2
Merkmale	• direkter Zugriff von Orientierung / Text auf Beispielwissen. • keine weitere Navigation in der Lerneinheit

Cluster 5/10; 11/28

Anzahl der Sequenzen	19
Prozentualer Anteil	3,6%
typische Abfolge	
Kennung (art_id)	2851 – (2885) – 2862
Wissensart (art)	Orient./T – (Handlung) - Erklärung
Anzahl der Elemente	3
Merkmale	• Ausgehend vom Orientierungswissen / Text über Handlung zu Erklärung, bzw. direkter Zugriff auf Erklärung ohne Zwischenschritt Handlung. • enthält keine weiteren Wissenseinheiten der Lerneinheit

Cluster 5/10; 12/28

Anzahl der Sequenzen	6
Prozentualer Anteil	1,1%
typische Abfolge	
Kennung (art_id)	3185 – 2851 - ...
Wissensart (art)	Orient./A – Orient./T - ....

Anzahl der Elemente	7
Merkmale	• Wechsel des Medientyps zu Beginn der Sequenz: danach jeweils Wechsel zwischen Handlung, Erklärung, Beispiel • kein Quellenwissen, kein Multiple Choice, kein True/False

Cluster 5/10; 13/28

Anzahl der Sequenzen	11
Prozentualer Anteil	2,1%
typische Abfolge	
Kennung (art_id)	2851 – 2885 – ... - 2851
Wissensart (art)	Orient./T – Handlung - ... - Orient./T
Anzahl der Elemente	5
Merkmale	• heterogenes Cluster • Sequenzbeginn von Orientierung / Text zu Handlung, danach 2 weitere Wissenseinheiten. Ende der Sequenz ist Orientierung. • kein Quellenwissen

Cluster 5/19; 14/28

Anzahl der Sequenzen	14
Prozentualer Anteil	2,7%
typische Abfolge	
Kennung (art_id)	2851 – 2885
Wissensart (art)	Orient./T - Handlung
Anzahl der Elemente	2
Merkmale	• homogenes Cluster • enthält keine weiteren Wissenseinheiten der Lerneinheit

Cluster 6/10; 15/28

Anzahl der Sequenzen	12
Prozentualer Anteil	2,3%
typische Abfolge	
Kennung (art_id)	2854
Wissensart (art)	[Grafik]
Anzahl der Elemente	1
Merkmale	• Dieser Aufruf ist technisch erzeugt und beruht nicht auf der Navigation eines Nutzers. Diese Wissenseinheit bezieht sich auf ein Grafik, die automatisch mit einer anderen Wissenseinheit geladen

	wurde und nur für einen kurzen Zeitraum in der Lernumgebung enthalten war. Diese Elemente werden bei der Analyse nicht berücksichtigt.

Cluster 6/10; 16/28

Anzahl der Sequenzen	10
Prozentualer Anteil	1,9%
typische Abfolge	
Kennung (art_id)	2854
Wissensart (art)	[Grafik]
Anzahl der Elemente	1
Merkmale	• vgl. Cluster 6/10; 15/28

Cluster 7/10; 17/28

Anzahl der Sequenzen	206
Prozentualer Anteil	39,2%
typische Abfolge	
Kennung (art_id)	2851
Wissensart (art)	Orient./T
Anzahl der Elemente	1
Merkmale	• alleiniger Aufruf der Wissenseinheit Orientierung / Text • enthält keine weiteren Wissenseinheiten der Lerneinheit

Cluster 8/10 / 18/28

Anzahl der Sequenzen	10
Prozentualer Anteil	1,9%
typische Abfolge	
Kennung (art_id)	2851 – 2885 – 2862 – 2863 – 2861 – 2889 – 2888
Wissensart (art)	Orient./T – Handlung – Erklärung – Beispiel – Quellen – True/False – Mutiple Choice
Anzahl der Elemente	7
Merkmale	• Navigation „von links nach rechts“ • kein Wechsel des Medientyps bei Orientierung

Cluster 8/10; 19/28

Anzahl der Sequenzen	6

Prozentualer Anteil	1,1%
typische Abfolge	
Kennung (art_id)	2851 – 2885 – 2862 – 2863 – 2861 – 2889 – 2888 - 2851
Wissensart (art)	Orient./T – Handlung – Erklärung – Beispiel – Quellen – True/False - Multiple Choice - Orient./T
Anzahl der Elemente	10
Merkmale	• vgl. Cluster 8 / 18: Sequenz wird nach diesem „Durchlauf von links nach rechts“ fortgesetzt; Endpunkt der Sequenz ist wieder der Ausgangspunkt (Orientierung) • kein Wechsel des Medientyps beim Orientierungswissen • ähnliche Abfolge wie Cluster 8/10 / 18/28, jedoch am Ende der Sequenz zunächst wieder zum Ausgangspunkt Orientierungswissen (2851) zurück

Cluster 8/10; 20/28

Anzahl der Sequenzen	7
Prozentualer Anteil	1,3%
typische Abfolge	
Kennung (art_id)	2851 – 2885 – 2862 – 2889 – 2888
Wissensart (art)	Orient./T – Handlung – Erklärung – True/False - Multiple Choice
Anzahl der Elemente	5
Merkmale	• Navigation „von links nach rechts“, jedoch ohne Quellenwissen. Sequenz endet mit dem Aufruf der Wissenseinheit Multiple Choice. Kein Zurückspringen zum Ausgangspunkt. • kein Quellenwissen • kein Wechsel des Medientyps beim Orientierungswissen

Cluster 8/10; 21/28

Anzahl der Sequenzen	4
Prozentualer Anteil	1,3%
typische Abfolge	
Kennung (art_id)	2851 – 2885 – 2862 – 2863 – 2889 – 2888 – 2899 - 2888
Wissensart (art)	Orient./T – Handlung – Erklärung – Beispiel – True/False – Multiple Choice – True/False - Multiple Choice
Anzahl der Elemente	1
Merkmale	• Navigation „von links nach rechts“, jedoch ohne Quellenwissen, dann wiederholter Aufruf von Multiple Choice und True/False

Cluster 8/10; 22/28

Anzahl der Sequenzen	2

Prozentualer Anteil	0,4%

Cluster 8/10; 23/28

Anzahl der Sequenzen	2
Prozentualer Anteil	0,4%
Anzahl der Elemente	14, 15

Cluster 8/10; 24/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,2%
Anzahl der Elemente	20

Cluster 8/10; 25/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,2%
Anzahl der Elemente	20

Cluster 9/10; 26/28

Anzahl der Sequenzen	17
Prozentualer Anteil	3,2%
typische Abfolge	
Kennung (art_id)	3185 – 2885 – 2862 – 2863 – 2861 – 2889 - 2888
Wissensart (art)	Orient./A – Handlung – Erklärung – Beispiel – Quellen - True/False – Multiple Choice – True/False - Multiple Choice
Anzahl der Elemente	7
Merkmale	• Navigation „von links nach rechts“. • Im Vergleich zu Cluster 8/10; 21/28 jedoch anderer Startpunkt (Orientierung / Animation anstatt Orientierung / Text).

Cluster 9/10; 27/28

Anzahl der Sequenzen	6
Prozentualer Anteil	1,1%
typische Abfolge	
Kennung (art_id)	3185 – 2885 – 2862 – 2863 – 2861 – 2889 – 2888 - ...
Wissensart (art)	Orient./A – Handlung – Erklärung – Beispiel – Quellen – True/False – Multiple Choice - ...

Anzahl der Elemente	11
Merkmale	• kein Wechsel des Medientyps bei Orientierungswissen. • Navigation „von links nach rechts“, danach fortgesetzte Navigation in Lerneinheit

Cluster 10/10 / 28/28

Anzahl der Sequenzen	6
Prozentualer Anteil	1,3%
typische Abfolge	
Kennung (art_id)	2851 – 3185 – 3863 – 2885 – 2862 – 2888 – 2889
Wissensart (art)	Orient./T – Orient./A – Beispiel – Handlung – Erklärung – Multiple Choice - True/False
Anzahl der Elemente	7
Merkmale	• Aufruf der Wissenseinheit im Rahmen der Kurses (Kursnavigation)

11.1.3 Lerneinheit 516: Median

Das folgende Dendogramm (vgl. Abb. 60) beschreibt den Agglomerationsprozess der Clusteranalyse für die Sequenzen der Lerneinheit „Median“ (516). Ausgangspunkt ist die Clusterlösung mit 28 Clustern, die in der linken Spalte dargestellt werden. Anhand des Dendogramms wird deutlich, welche Cluster in den folgenden Schritten fusioniert werden. Darüber hinaus verdeutlichen die Kantenlängen auf der x-Achse die Zunahme des Fusionskoeffizienten bei fortschreitender Fusionierung der Cluster.

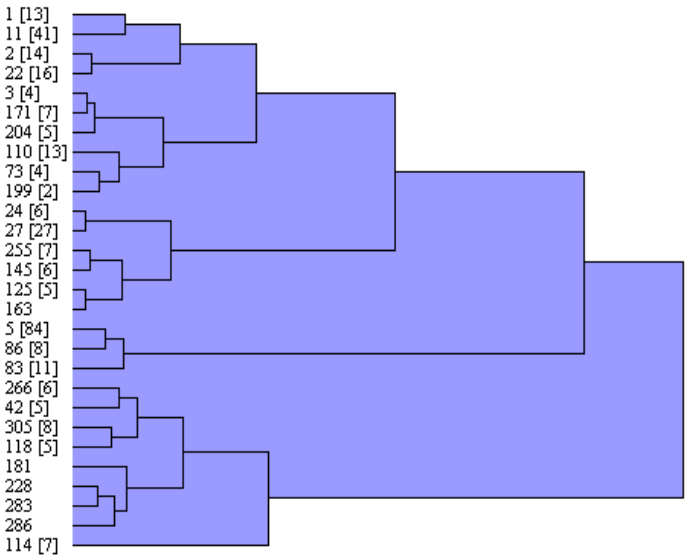

Abbildung 60: Clusterlösung der Lerneinheit „Median“ (516).

Die Zunahme des Fusionskoeffizienten bei der Fusionierung von Clustern wird anhand folgender Grafik deutlich (vgl. Abb. 62).

Für die Daten der Lerneinheit „Median“ beträgt der t-Wert für die Clusterlösung mit 28 Clustern 6,77 und für die Clusterlösung mit 10 Clustern 30,96. Statistisch signifikant auf einem 5% Niveau sind alle Clusterlösungen mit weniger als 43 Clustern.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	54	17,5	17,5	17,5
	2	30	9,7	9,7	27,2
	3	16	5,2	5,2	32,4
	4	19	6,1	6,1	38,5
	5	33	10,7	10,7	49,2
	6	19	6,1	6,1	55,3
	7	103	33,3	33,3	88,7
	8	24	7,8	7,8	96,4
	9	4	1,3	1,3	97,7
	10	7	2,3	2,3	100,0
	Gesamt	309	100,0	100,0	

Abbildung 61: Häufigkeitsverteilung der Ward-Clusterlösung mit 10 Clustern (Lerneinheit 516, „Median“).

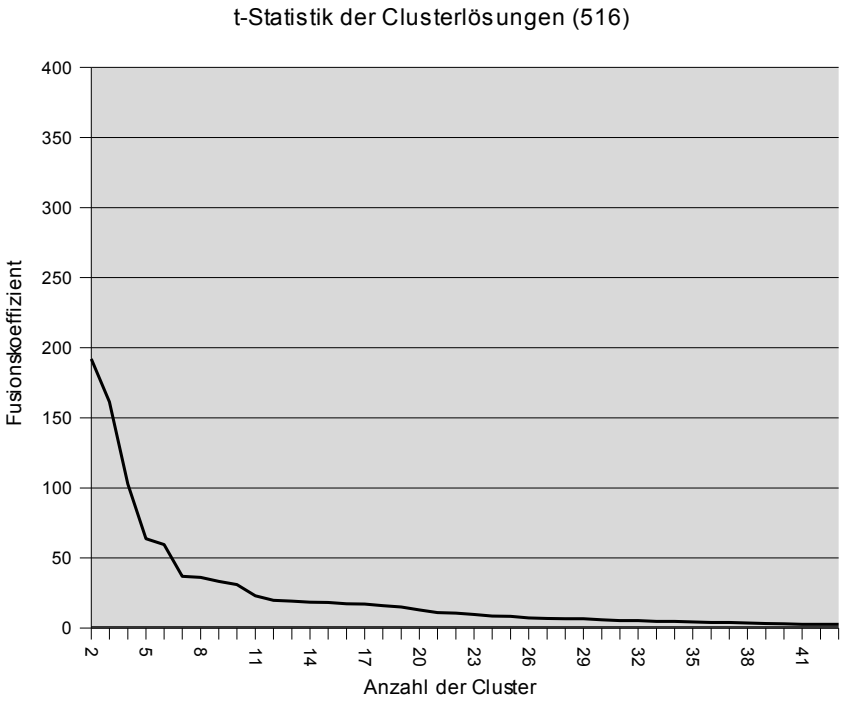

Abbildung 62: t-Statistik der Clusterlösungen, „Median“ (516)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	13	4,2	4,2	4,2
	2	41	13,3	13,3	17,5
	3	14	4,5	4,5	22,0
	4	16	5,2	5,2	27,2
	5	4	1,3	1,3	28,5
	6	7	2,3	2,3	30,7
	7	5	1,6	1,6	32,4
	8	13	4,2	4,2	36,6
	9	4	1,3	1,3	37,9
	10	2	,6	,6	38,5
	11	6	1,9	1,9	40,5
	12	27	8,7	8,7	49,2
	13	7	2,3	2,3	51,5
	14	6	1,9	1,9	53,4
	15	5	1,6	1,6	55,0
	16	1	,3	,3	55,3
	17	84	27,2	27,2	82,5
	18	8	2,6	2,6	85,1
	19	11	3,6	3,6	88,7
	20	6	1,9	1,9	90,6
	21	5	1,6	1,6	92,2
	22	8	2,6	2,6	94,8
	23	5	1,6	1,6	96,4
	24	1	,3	,3	96,8
	25	1	,3	,3	97,1
	26	1	,3	,3	97,4
	27	1	,3	,3	97,7
	28	7	2,3	2,3	100,0
	Gesamt	309	100,0	100,0	

Abbildung 63: Häufigkeitsverteilung der Ward-Clusterlösung mit 28 Clustern (Lerneinheit 516, „Median“)

Für einen Überblick über das Ergebnis der Clusteranalyse wird die Häufigkeitsverteilung für die Clusterlösung mit 10 Clustern (vgl. Abb. 61) und mit 28 Clustern (vgl. Abb. 63) tabellarisch dargestellt. Diese tabellarische Darstellung dient als Ausgangspunkt für die folgende formale Beschreibung der Sequenzen der Cluster.

Die Anzahl der analysierten Sequenzen in dieser Lerneinheit (516) beträgt insgesamt 212, mit insgesamt 604 Wissenseinheiten.

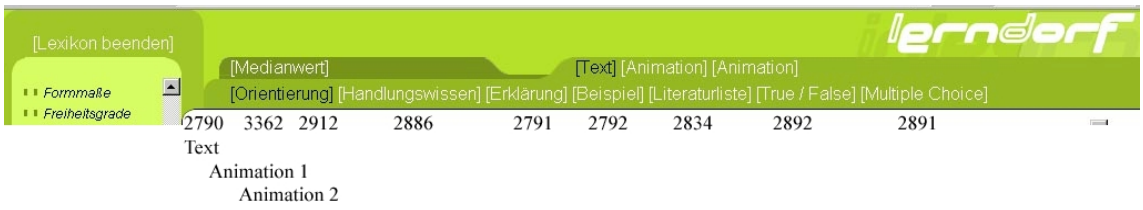

Abbildung 64: Wissenseinheit „Median“ (516), Kennung – Wissensart.

Im Folgenden werden die Clusterlösungen tabellarisch dargestellt und formal beschrieben:

Cluster 1/10; 1/28

Anzahl der Sequenzen	13
Prozentualer Anteil	4,2%
typische Abfolge	
Kennung (art_id)	3362 – 2886
Wissensart (art)	Orient./A2 - Handlung
Anzahl der Elemente	2
Merkmale	• Navigation „von links nach rechts“ • keine weiteren Wissenseinheiten der Lerneinheit • homogenes Cluster

Cluster 1/10; 2/28

Anzahl der Sequenzen	41
Prozentualer Anteil	13,5%
typische Abfolge	
Kennung (art_id)	3362
Wissensart (art)	Orient./A1
Anzahl der Elemente	1
Merkmale	• in diesem Cluster befinden sich 3 Sequenzen mit der Abfolge Orientierung(3362) – Beispiel (2792);

	sowie 2 Sequenzen mit der Abfolge Orientierung (3362) – Quellen (2834).

Cluster 2/10; 3/28 und 4/28:

Anzahl der Sequenzen	30
Prozentualer Anteil	9,7%
typische Abfolge	
Kennung (art_id)	2790 – 3363 – (2912)
Wissensart (art)	Orient./T – Orient./A1 - (Orient./A2)
Anzahl der Elemente	2 (3)
Merkmale	• Wechsel der Medientypen des Orientierungswissens von Orientierung / Text zu Orientierung / Animation 1, wahlweise weiter zu Orientierung / Animation 2 • keine weiteren Wissenseinheiten

Cluster 3/10; 5/28

Anzahl der Sequenzen	4
Prozentualer Anteil	1,3%
typische Abfolge	
Kennung (art_id)	2892
Wissensart (art)	True/False
Anzahl der Elemente	1
Merkmale	• Aufruf der Wissenseinheit True/False innerhalb der Kursnavigation

Cluster 3/10; 6/28

Anzahl der Sequenzen	7
Prozentualer Anteil	2,3%
typische Abfolge	
Kennung (art_id)	3362 – 2891
Wissensart (art)	Orient./A1 - Multiple Choice
Anzahl der Elemente	2
Merkmale	• direkter Zugriff auf Aufgabe Multiple Choice • keine weiteren Wissenseinheiten

Cluster 3/10; 7/28

Anzahl der Sequenzen	5
Prozentualer Anteil	1,6%

typische Abfolge	
Kennung (art_id)	3362 – 2791 – 2891 – 2892
Wissensart (art)	Orient./A1 – Erklärung – Multiple Choice - True/False
Anzahl der Elemente	4
Merkmale	• direkter Zugriff auf Wissenseinheit Erklärung und Aufgabe (Multiple Choice, True/False) • keine Navigation „von links nach rechts“

Cluster 4/10; 8/28

Anzahl der Sequenzen	13
Prozentualer Anteil	4,2%
typische Abfolge	
Kennung (art_id)	2790 - 2791
Wissensart (art)	Orient./T - Erklärung
Anzahl der Elemente	2
Merkmale	• direkter Zugriff auf die Wissenseinheit Erklärung • keine „von links nach rechts“ Navigation • keine weitere Navigation in der Wissenseinheit: kein Beispiel, Quellen, Multiple Choice, kein True/False

Cluster 4/10; 9/28

Anzahl der Sequenzen	4
Prozentualer Anteil	1,3%
typische Abfolge	
Kennung (art_id)	2790 – 2792 - ...
Wissensart (art)	Orient./T – Beispiel - ...
Anzahl der Elemente	6
Merkmale	• heterogenes Cluster • kein Wechsel des Medientyps beim Orientierungswissen • direkte Navigation von Orientierungswissen / Text zu Beispiel; danach weitere 4 Wissenseinheiten • kein Quellenwissen

Cluster 4/10; 10/28

Anzahl der Sequenzen	2
Prozentualer Anteil	0,6%

Cluster 5/10; 11/28

Anzahl der Sequenzen	6
Prozentualer Anteil	1,9%
typische Abfolge	
Kennung (art_id)	2912 - 2791
Wissensart (art)	Orient./A2 - Erklärung
Anzahl der Elemente	2
Merkmale	• homogenes Cluster • kein Wechsel des Medientyps beim Orientierungswissen • direkter Zugriff von Orientierungswissen auf Erklärungswissen, keine weitergehende Navigation in der Lerneinheit.

Cluster 5/10; 12/28

Anzahl der Sequenzen	27
Prozentualer Anteil	8,7%
typische Abfolge	
Kennung (art_id)	2912
Wissensart (art)	Orient./A2
Anzahl der Elemente	1
Merkmale	• Aufruf der Wissenseinheit Orientierung / Animation 2, keine weitere Navigation innerhalb der Lerneinheit

Cluster 6/10; 13/28

Anzahl der Sequenzen	7
Prozentualer Anteil	2,3%
typische Abfolge	
Kennung (art_id)	2912 – 2886 - 2791
Wissensart (art)	Orient./A2 – Handlung - Erklärung
Anzahl der Elemente	3
Merkmale	• Navigation „von links nach rechts“ • kein Quellenwissen, True/False, Multiple Choice • kein Wechsel zwischen Medientypen des Orientierungswissens

Cluster 6/10; 14/28

Anzahl der Sequenzen	6

Prozentualer Anteil	1,9%
typische Abfolge	
Kennung (art_id)	2912 – 2792
Wissensart (art)	Orient./A2 - Beispiel
Anzahl der Elemente	2
Merkmale	• direkter Zugriff auf die Wissenseinheit Beispiel • keine „von links nach rechts“ Navigation

Cluster 6/10; 15/28

Anzahl der Sequenzen	5
Prozentualer Anteil	1,6%
typische Abfolge	
Kennung (art_id)	2912 – 2791 - ... - 2790
Wissensart (art)	Orient./A2 – Erklärung - ... - Orientierung
Anzahl der Elemente	5
Merkmale	• von Orientierung / Animation2 direkter Zugriff auf Erklärung, danach zwei weitere Wissenseinheiten. Sequenz endet mit Orientierung • kein Quellenwissen, kein True/False, kein Multiple Choice • keine „von links nach rechts“ Navigation • heterogenes Cluster

Cluster 6/10; 16/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,3%

Cluster 7/10; 17/28

Anzahl der Sequenzen	84
Prozentualer Anteil	27,2%
typische Abfolge	
Kennung (art_id)	2790
Wissensart (art)	Orient./T
Anzahl der Elemente	1
Merkmale	• homogenes Cluster • alleiniger Aufruf der Wissenseinheit Orientierung / Text

Cluster 7/10; 18/28

Anzahl der Sequenzen	8
Prozentualer Anteil	2,6%
typische Abfolge	
Kennung (art_id)	2790 – 2792
Wissensart (art)	Orient./T - Beispiel
Anzahl der Elemente	2
Merkmale	• homogenes Cluster • nach Orientierung / Text direkter Zugriff auf Beispiel • keine „von links nach rechts“ Navigation • kein Wechsel des Medientyps innerhalb des Orientierungswissens • keine weiteren Wissenseinheiten der Lerneinheit

Cluster 7/10; 19/28

Anzahl der Sequenzen	11
Prozentualer Anteil	3,6%
typische Abfolge	
Kennung (art_id)	2790 - 2886
Wissensart (art)	Orient./T - Handlung
Anzahl der Elemente	2
Merkmale	• homogenes Cluster • nach Orientierungswissen / Text direkter Zugriff auf Handlung • keine „von links nach rechts“ Navigation • kein Medienwechsel innerhalb des Orientierungswissens • keine weiteren Wissenseinheiten der Lerneinheit

Cluster 8/10; 20/28

Anzahl der Sequenzen	6
Prozentualer Anteil	1,9%
typische Abfolge	
Kennung (art_id)	2790 – 2886 – 2791 – 2792 – 2834 – 2892 - 2991 - 2790
Wissensart (art)	Orient/T – Handlung – Erklärung – Beispiel – Literatur – True/False – Multiple Choice - Orient/T
Anzahl der Elemente	8
Merkmale	• Navigation „von links nach rechts“, danach zum Ausgangspunkt (Orientierungswissen) zurück. • kein Wechsel des Medientyps innerhalb des Orientierungswissens

	• homogenes Cluster

Cluster 8/10; 21/28

Anzahl der Sequenzen	5
Prozentualer Anteil	1,6%
typische Abfolge	
Kennung (art_id)	2912 – 2886 - 2791 – 2792 – 2892 - 2891
Wissensart (art)	Orient/A2 – Handlung - Erklärung – Beispiel – True/False - Multiple Choice
Anzahl der Elemente	6
Merkmale	• Navigation „von links nach rechts“ • kein Wechsel des Medientyps innerhalb des Orientierungswissens • im Vergleich zu Cluster 8/10; 21/28 ohne Quellenwissen (2834) und am Ende der Sequenz kein Sprung zurück zum Ausgangspunkt

Cluster 8/10; 22/28

Anzahl der Sequenzen	8
Prozentualer Anteil	2,6%
typische Abfolge	
Kennung (art_id)	2790 – 2886 – 2791 – 2792 – (2886 / 2892)
Wissensart (art)	Orient./T – Handlung – Erklärung – Beispiel – (Handlung, Multiple Choice)
Anzahl der Elemente	5
Merkmale	• ausgehend vom Orientierungswissen / Text Navigation „von links nach rechts; danach Wahl einer weiteren Wissenseinheit: Handlung bzw. Multiple Choice • kein Quellenwissen

Cluster 8/10; 23/28

Anzahl der Sequenzen	5
Prozentualer Anteil	1,6%
typische Abfolge	
Kennung (art_id)	3362 – 2886 – 2791 – 2792 - 2997
Wissensart (art)	Orient./A1 – Handlung – Erklärung – Beispiel – Aufgabe/T
Anzahl der Elemente	5
Merkmale	• Navigation „von links nach rechts“ • kein Wechsel des Medientyps beim Orientierungswissen • kein Quellenwissen

Cluster 9/10; 24/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,3%

Cluster 9/10; 25/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,3%

Cluster 9/10; 26/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,3%

Cluster 9/10; 27/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,3%

Cluster 10/10; 28/28

Anzahl der Sequenzen	7
Prozentualer Anteil	2,3%
typische Abfolge	
Kennung (art_id)	2790 – 3362 – 2912 – 2792 – 2997 – 2886 – 2791 – 2891 – 2892
Wissensart (art)	Orient./T – Orient./A1 – Orient./A2 – Beispiel – Aufgabe/T – Handlung – Erklärung – Multiple Choice - True/False
Anzahl der Elemente	9
Merkmale	• homogenes Cluster • Abfolge entspricht der Kursnavigation

11.1.4 Lerneinheit 517: Modus

Das folgende Dendogramm (vgl. Abb. 65) beschreibt den Agglomerationsprozess der Clusteranalyse für die Sequenzen der Lerneinheit „Modus“ (517). Ausgangspunkt ist die Clusterlösung mit 28 Clustern, die in der linken Spalte dargestellt werden. Anhand des Dendogramms wird deutlich, welche Cluster in den folgenden Schritten fusioniert werden. Darüber hinaus verdeutlichen die Länge der Linien auf der x-Achse die Zunahme des Fusionskoeffizienten bei fortschreitender Fusionierung der Cluster.

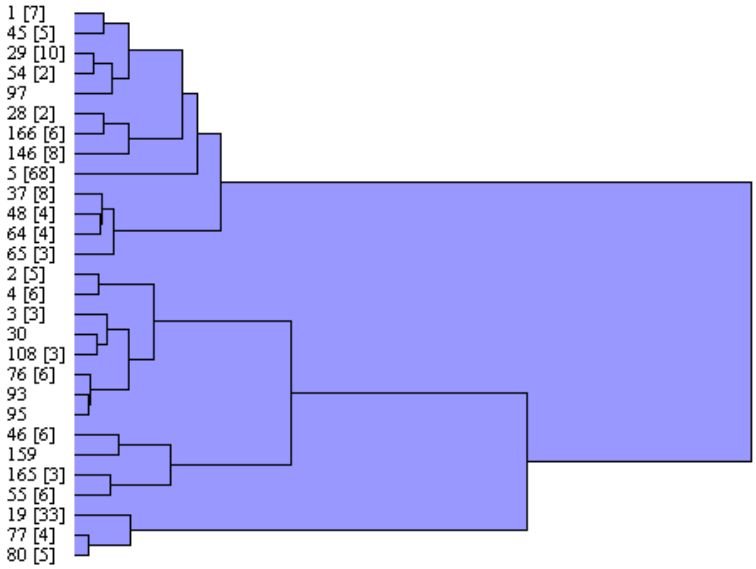

Abbildung 65: Clusterlösung der Lerneinheit „Modus“ (517).

Die Zunahme des Fusionskoeffizienten bei der Fusionierung von Clustern wird anhand folgender Grafik deutlich (vgl. Abb. 66): Für die Daten der Lerneinheit „Arithmetisches Mittel“ beträgt der t-Wert für die Clusterlösung mit 28 Clustern 2,73 und für die Clusterlösung mit 10 Clustern 12,37. Statistisch signifikant auf einem 5% Niveau sind alle Clusterlösungen mit weniger als 29 Clustern.

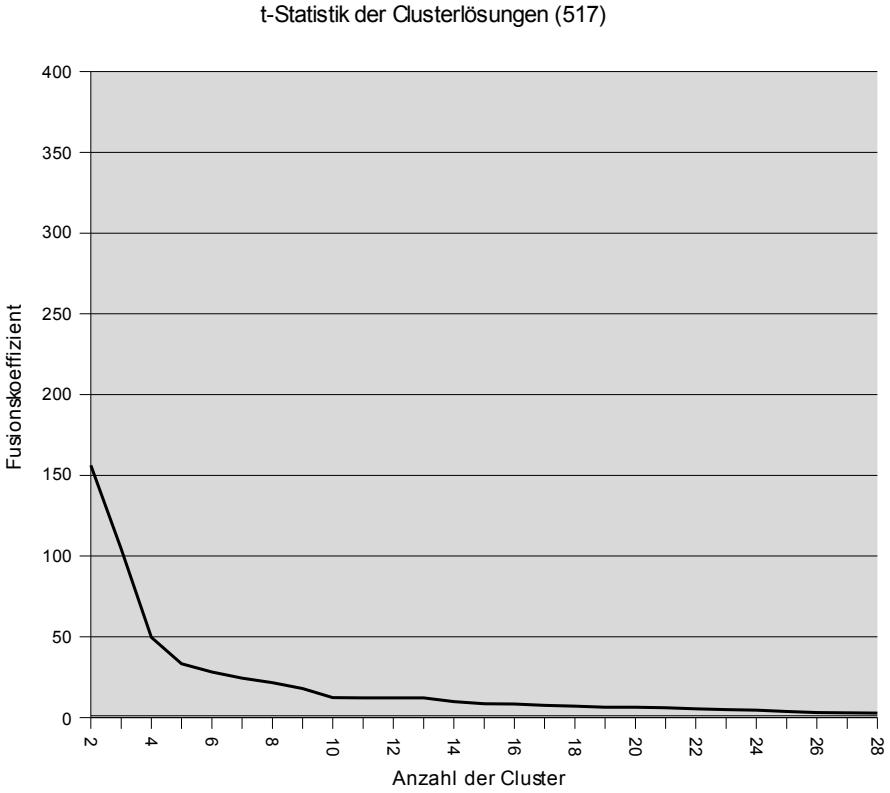

Abbildung 66: t-Statistik der Clusterlösungen, „Modus“ (517).

Für einen Überblick über das Ergebnis der Clusteranalyse wird die Häufigkeitsverteilung für die Clusterlösung mit 10 Clustern (vgl. Abb. 67) und mit 28 Clustern (vgl. Abb. 68) tabellarisch dargestellt. Diese tabellarische Darstellung dient als Ausgangspunkt für die folgende formale Beschreibung der Sequenzen der Cluster.

cg_scostdefault_ward_10c

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	25	11,8	11,8	11,8
	2	16	7,5	7,5	19,3
	3	68	32,1	32,1	51,4
	4	19	9,0	9,0	60,4
	5	11	5,2	5,2	65,6
	6	15	7,1	7,1	72,6
	7	7	3,3	3,3	75,9
	8	9	4,2	4,2	80,2
	9	33	15,6	15,6	95,8
	10	9	4,2	4,2	100,0
	Gesamt	212	100,0	100,0	

Abbildung 67: Häufigkeitsverteilung der Ward-Clusterlösung mit 10 Clustern (Lerneinheit 517, „Modus“).

cg_scostdefault_ward_28c

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	7	3,3	3,3	3,3
	2	5	2,4	2,4	5,7
	3	10	4,7	4,7	10,4
	4	2	,9	,9	11,3
	5	1	,5	,5	11,8
	6	2	,9	,9	12,7
	7	6	2,8	2,8	15,6
	8	8	3,8	3,8	19,3
	9	68	32,1	32,1	51,4
	10	8	3,8	3,8	55,2
	11	4	1,9	1,9	57,1
	12	4	1,9	1,9	59,0
	13	3	1,4	1,4	60,4
	14	5	2,4	2,4	62,7
	15	6	2,8	2,8	65,6
	16	3	1,4	1,4	67,0
	17	1	,5	,5	67,5
	18	3	1,4	1,4	68,9
	19	6	2,8	2,8	71,7
	20	1	,5	,5	72,2
	21	1	,5	,5	72,6
	22	6	2,8	2,8	75,5
	23	1	,5	,5	75,9
	24	3	1,4	1,4	77,4
	25	6	2,8	2,8	80,2
	26	33	15,6	15,6	95,8
	27	4	1,9	1,9	97,6
	28	5	2,4	2,4	100,0
	Gesamt	212	100,0	100,0	

Abbildung 68: Häufigkeitsverteilung der Ward-Clusterlösung mit 28 Clustern (Lerneinheit 517, „Modus“).

Die Anzahl der analysierten Sequenzen in dieser Lerneinheit (517) beträgt insgesamt 309, mit insgesamt 896 Wissenseinheiten.

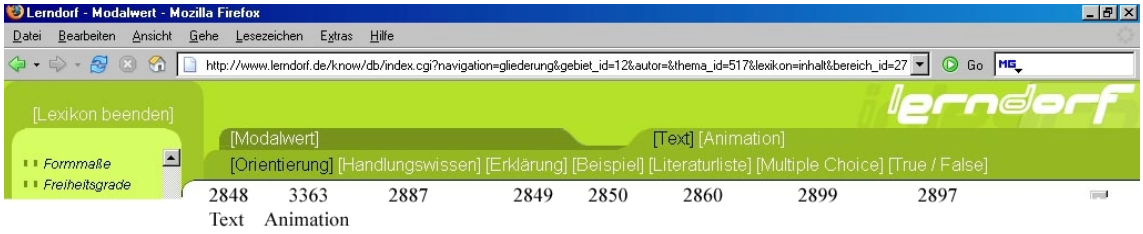

Abbildung 69: Wissenseinheit „Modus“ (517), Kennung – Wissensart.

Im Folgenden werden die Clusterlösungen tabellarisch dargestellt und formal beschrieben:

Cluster 1/10; 1/28

Anzahl der Sequenzen	7
Prozentualer Anteil	3,3%
typische Abfolge	
Kennung (art_id)	2848 – 3363

Wissensart (art)	Orient./T - Orient./A
Anzahl der Elemente	2
Merkmale	• homogenes Cluster • Wechsel des Medientyps beim Orientierungswissen von Text zu Animation • keine weitere Navigation innerhalb der Lerneinheit

Cluster 1/10; 2/28

Anzahl der Sequenzen	5
Prozentualer Anteil	2,4%
typische Abfolge	
Kennung (art_id)	2848 – 2850
Wissensart (art)	Orient./T - Beispiel
Anzahl der Elemente	2
Merkmale	• direkter Zugriff von Orientierung auf Beispiel • keine weitere Navigation innerhalb der Lerneinheit

Cluster 1/10; 3/28

Anzahl der Sequenzen	10
Prozentualer Anteil	4,7%
typische Abfolge	
Kennung (art_id)	3363 – 2848 - (3363)
Wissensart (art)	Orient./A – Orient./T - (Orient./A)
Anzahl der Elemente	2 (3)
Merkmale	• Wechsel des Medientyps beim Orientierungswissen von Animation zu Text (und zurück zu Text) • keine weitere Navigation innerhalb der Lerneinheit

Cluster 1/10; 4/28

Anzahl der Sequenzen	2
Prozentualer Anteil	0,9%
typische Abfolge	
Kennung (art_id)	3363 – 2887 – 3363 – 2848
Wissensart (art)	Orient./A – Handlung – Orient./A - Orient./T
Anzahl der Elemente	4
Merkmale	• Navigation von Orientierung / Animation zu Handlung, dann zurück zu Orientierung / Animation und Wechsel des Medientyps zu Orientierung / Text

Cluster 1/10; 5/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,5%

Cluster 2/10; 6/28

Anzahl der Sequenzen	2
Prozentualer Anteil	0,9%
typische Abfolge	
Kennung (art_id)	2897 – 2899 - 2849 – 2887
Wissensart (art)	True/False – Multiple Choice – Erklärung – Beispiel
Anzahl der Elemente	4
Merkmale	• Kursnavigation: „zurück“-Navigation im Kurs über „Player“ von True/False bis zu Beispiel

Cluster 2/10; 7/28

Anzahl der Sequenzen	6
Prozentualer Anteil	2,8%
typische Abfolge	
Kennung (art_id)	2848 – 2897
Wissensart (art)	Orient./T - True/False
Anzahl der Elemente	2
Merkmale	• direkter Zugriff von Orient./T auf True/False • keine weitere Navigation innerhalb der Lerneinheit

Cluster 3/10; 8/28

Anzahl der Sequenzen	8
Prozentualer Anteil	3,8%
typische Abfolge	
Kennung (art_id)	2848 – 2899 – 2897
Wissensart (art)	Orient./T – Multiple Choice - True/False
Anzahl der Elemente	3
Merkmale	• direkter Zugriff von Orient./T auf Multiple Choice, danach True/False • keine weitere Navigation innerhalb der Lerneinheit

Cluster 3/10; 9/28

Anzahl der Sequenzen	68
Prozentualer Anteil	32,1%
typische Abfolge	
Kennung (art_id)	2848
Wissensart (art)	Orient./T
Anzahl der Elemente	1
Merkmale	• einmaliger Aufruf der Wissenseinheit Orientierung / Text • keine weitere Navigation innerhalb der Lerneinheit

Cluster 4/10; 10/28

Anzahl der Sequenzen	8
Prozentualer Anteil	3,6%
typische Abfolge	
Kennung (art_id)	2848 - 2887
Wissensart (art)	Orient./T - Handlungswissen
Anzahl der Elemente	2
Merkmale	• homogenes Cluster • kein Wechsel des Medientyps innerhalb des Orientierungswissens

Cluster 4/10; 11/28

Anzahl der Sequenzen	4
Prozentualer Anteil	1,9%
typische Abfolge	
Kennung (art_id)	2848 – 2887 – 2849 - 2850
Wissensart (art)	Orient./T – Handlung – Erklärung – Beispiel
Anzahl der Elemente	4
Merkmale	• homogenes Cluster • Navigation „von links nach rechts“ • keine weitere Navigation innerhalb der Lerneinheit

Cluster 4/10; 12/28

Anzahl der Sequenzen	4
Prozentualer Anteil	1,9%
typische Abfolge	

Kennung (art_id)	2848 – 2849
Wissensart (art)	Orient./T - Erklärung
Anzahl der Elemente	2
Merkmale	• direkter Zugriff von Orientierungswissen / Text auf Erklärung

Cluster 4/10; 13/28

Anzahl der Sequenzen	3
Prozentualer Anteil	1,4%
typische Abfolge	
Kennung (art_id)	2848 – 2849 – 2887 – 2849 – 2887
Wissensart (art)	OW/T – Erklärung – Handlung -– Erklärung – Handlung
Anzahl der Elemente	5
Merkmale	• Navigation von Orientierungswissen zu Erklärung, danach Handlung. Erneut Handlung (2887) und Erklärung (2849) • vgl. Cluster 4 / 12; aber Navigation wird fortgesetzt • kein Quellenwissen, Multiple Choice, True/False

Cluster 5/10; 14/28

Anzahl der Sequenzen	5
Prozentualer Anteil	2,4%
typische Abfolge	
Kennung (art_id)	3363 – 2848 – 2887 – 2849
Wissensart (art)	Orient./A – Orient./T – Handlung – Erklärung
Anzahl der Elemente	4
Merkmale	• Wechsel des Medientyps von Orientierung / Animation zu Orientierung / Text. Dann Handlung und Erklärung. • keine weitere Navigation in der Lerneinheit

Cluster 5/10; 15/28

Anzahl der Sequenzen	6
Prozentualer Anteil	2,8%
typische Abfolge	
Kennung (art_id)	3363 - 2849
Wissensart (art)	Orient./A – Erklärung
Anzahl der Elemente	2
Merkmale	• homogenes Cluster

	• direkter Zugriff von Orientierung / Animation auf Erklärung • keine weitere Navigation innerhalb der Lerneinheit

Cluster 6/10; 16/28

Anzahl der Sequenzen	3
Prozentualer Anteil	1,4%
typische Abfolge	
Kennung (art_id)	3363 – 2849 – ...
Wissensart (art)	Orient/A – Handlung – ...
Anzahl der Elemente	5
Merkmale	• Beginn der Sequenz analog Cluster 5/10; 15/28, aber die Navigation wird in der Lerneinheit fortgesetzt • kein Quellenwissen, Multiple Choice, True/False

Cluster 6/10; 17/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,5%

Cluster 6/10; 18/28

Anzahl der Sequenzen	3
Prozentualer Anteil	1,4%
typische Abfolge	
Kennung (art_id)	3363 – 2850 – 2887 – 2849 - 2850
Wissensart (art)	Orient./A – Beispiel – Handlung – Erklärung – Beispiel
Anzahl der Elemente	5
Merkmale	• direkter Zugriff von Orientierung / Animation auf Beispiel, danach Handlung und Erklärung, dann wieder Beispiel. • kein Quellenwissen , Multiple Choice, True/False

Cluster 6/10; 19/28

Anzahl der Sequenzen	6
Prozentualer Anteil	2,8%
typische Abfolge	
Kennung (art_id)	3363 – 2887 – 2849 – 2850 – 2899 – 2897
Wissensart (art)	Orient./A – Handlung – Erklärung – Beispiel – Multiple Choice - True/False

Anzahl der Elemente	6
Merkmale	• homogenes Cluster • Navigation „von links nach rechts“, wobei Quellenwissen (2860) übersprungen wird • kein Wechsel des Medientyps beim Orientierungswissen

Cluster 6/10; 20/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,5%

Cluster 6 /10; 21/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,5%

Cluster 7/10; 22/28

Anzahl der Sequenzen	6
Prozentualer Anteil	2,8%
typische Abfolge	
Kennung (art_id)	2848– 2887 – 2849 – 2850 – (2860) - 2899 – 2897
Wissensart (art)	Orient./T – Handlung – Erklärung – Beispiel – (Quellenwissen) - Multiple Choice - True/False
Anzahl der Elemente	7
Merkmale	• Navigation „von links nach rechts“, mit und ohne Aufruf des Quellenwissens • vgl. Cluster 6 / 19, jedoch anderer Startpunkt und insgesamt längere Sequenz • heterogenes Cluster

Cluster 7/10; 23/28

Anzahl der Sequenzen	1
Prozentualer Anteil	0,5%

Cluster 8/10; 24/28

Anzahl der Sequenzen	3
Prozentualer Anteil	1,4%
typische Abfolge	
Kennung (art_id)	2848 – 3363 – 2850 – 2887 – 2849 – 2887 – 2849 - 2899 - 2897
Wissensart (art)	Orient./T – Orient./A – Beispiel – Handlung – Erklärung – Handlung – Erklärung – Multiple Choice -

	True/False
Anzahl der Elemente	9
Merkmale	• homogenes Cluster • Kursnavigation, jedoch mit einem Schritt zurück vom Erklärungswissen zum Handlungswissen, dann weiter in der Kursabfolge

Cluster 8/10; 25/28

Anzahl der Sequenzen	6
Prozentualer Anteil	2,8%
typische Abfolge	
Kennung (art_id)	2848 – 3363 – 2850 – 2887 – 2849 – 2899 – 2897
Wissensart (art)	Orient./T – Orient./A – Beispiel – Handlung – Erklärung – Multiple Choice - True/False
Anzahl der Elemente	7
Merkmale	• homogenes Cluster • Kursnavigation, ohne Schritt zurück (vgl. Cluster 9/10; 25/28)

Cluster 9 / 10; 26/28

Anzahl der Sequenzen	33
Prozentualer Anteil	15,6%
typische Abfolge	
Kennung (art_id)	3363
Wissensart (art)	Orient./A
Anzahl der Elemente	1
Merkmale	• Aufruf der Wissenseinheit Orientierung / Animation • keine weiter Navigation innerhalb der Lerneinheit

Cluster 10/10; 27/28

Anzahl der Sequenzen	4
Prozentualer Anteil	1,9%
typische Abfolge	
Kennung (art_id)	3363 – 2887 - 2897
Wissensart (art)	Orient./A – Handlung - True/False
Anzahl der Elemente	3
Merkmale	• Navigation vom Orientierungswissen / Animation zu Handlung, dann zu True/False • keine weitere Navigation innerhalb der Lerneinheit

Cluster 10/10; 28/28

Anzahl der Sequenzen	5
Prozentualer Anteil	2,4%
typische Abfolge	
Kennung (art_id)	3363 - 2887
Wissensart (art)	Orient./A - Handlung
Anzahl der Elemente	2
Merkmale	• homogenes Cluster • Navigation „von links nach rechts“ von Orientierung / Animation zu Handlung • kein Wechsel des Medientyps beim Orientierungswissen • keine weitere Navigation innerhalb der Lerneinheit

11.1.5 Navigationssequenzen im Überblick

In diesem Abschnitt werden ausgehend von den oben detailliert dargestellten Clusterlösungen die typischen Abfolgen der Navigationssequenzen in den spezifischen Lerneinheiten zusammenfassend grafisch dargestellt. In der linken Spalte befindet sich die Kennzeichnung des Clusters; im rechten Teil der Grafik wird die typische Abfolge visualisiert: Schwarze Quadrate verweisen dabei auf besuchte Wissenseinheiten, leere Quadrate verweisen auf ausgelassene bzw. nicht-aufgerufene Wissenseinheiten. In heterogenen Sequenzen enthaltene Variationen werden durch einen leeren Kreis angedeutet (vgl. 515: 11/28).

Ein heterogener Navigationsverlauf, bei dem die Abfolge nicht zusammenfassend beschreibbar und darstellbar ist, wird durch ein aufgezogenes Dreieck angedeutet, innerhalb dessen die weitere Navigation stattfindet: Die Visualisierung der Clusterlösung 8/29 der Lerneinheit 513 (vgl. Abb. 70) bedeutet daher, dass die Wissenseinheiten in der Abfolge Orientierung / Animation, Handlung und Erklärung aufgerufen wurden, gefolgt von einem heterogenen, mehrmaligen Wechsel zwischen Wissenseinheiten. Insgesamt besteht die typische Abfolge aus 12 Wissenseinheiten, Endpunkt der Sequenz ist das Orientierungswissen.

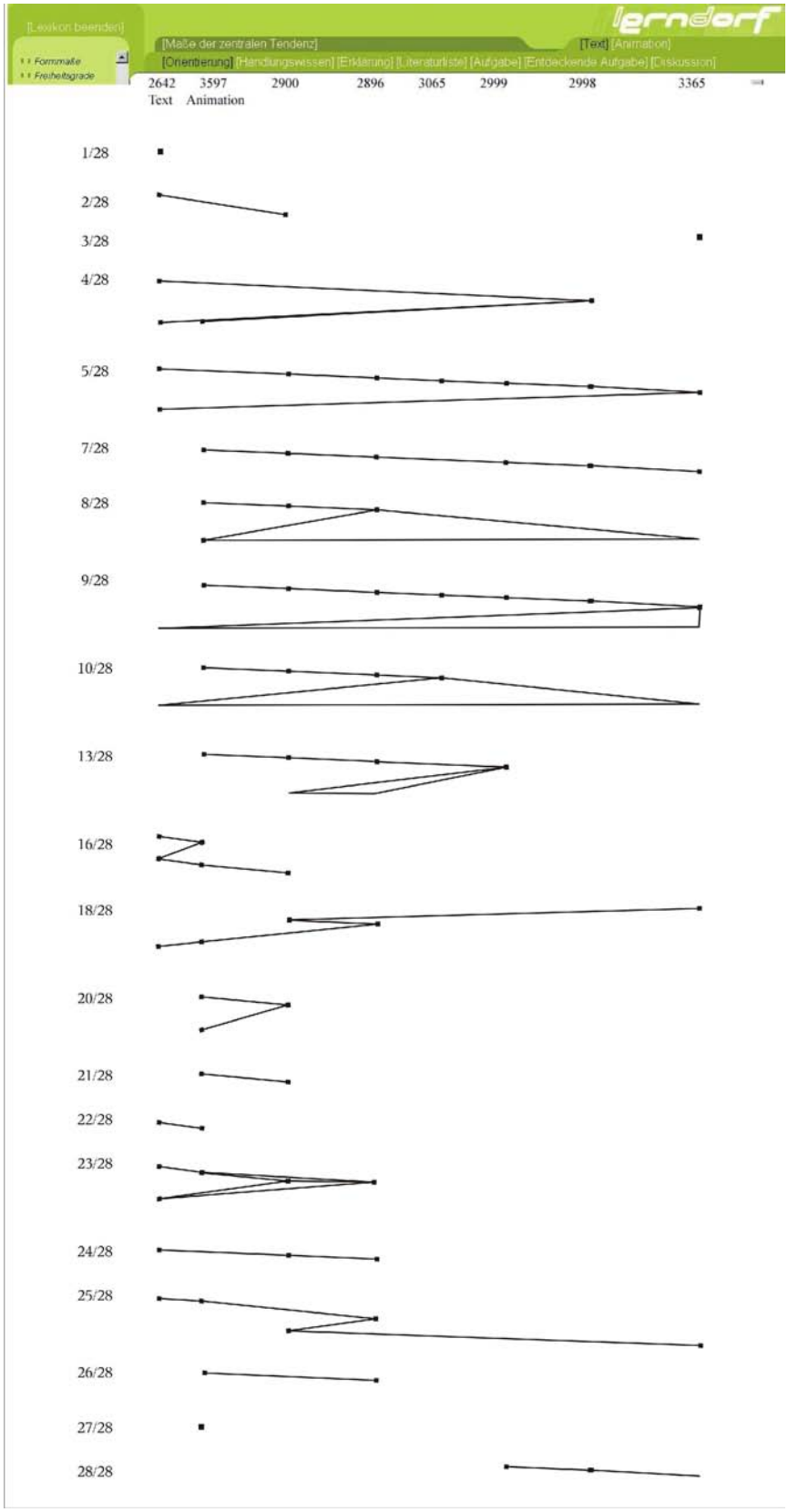

Abbildung 70: Navigationsmuster "Maße der zentralen Tendenz" (513).

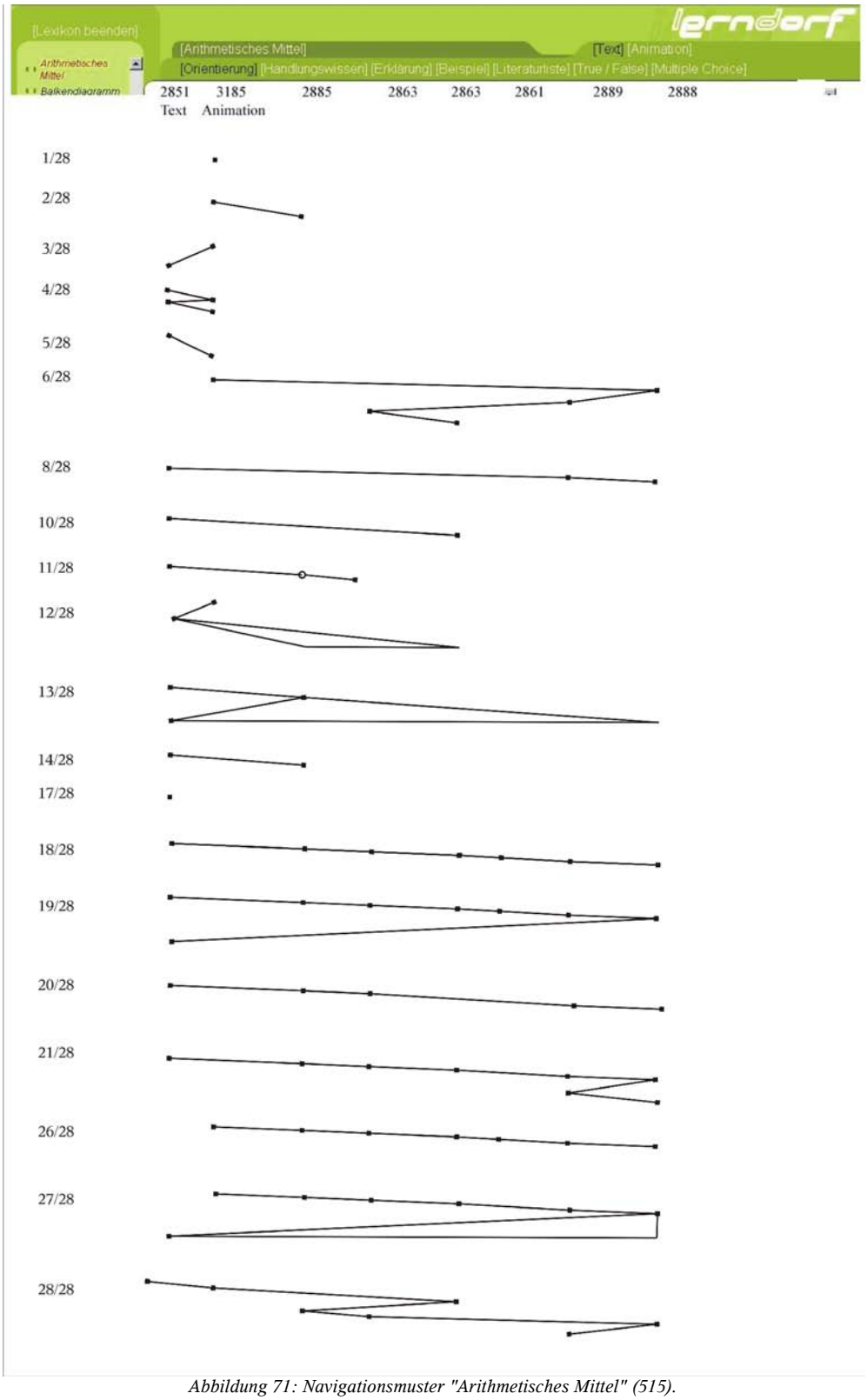

Abbildung 71: Navigationsmuster "Arithmetisches Mittel" (515).

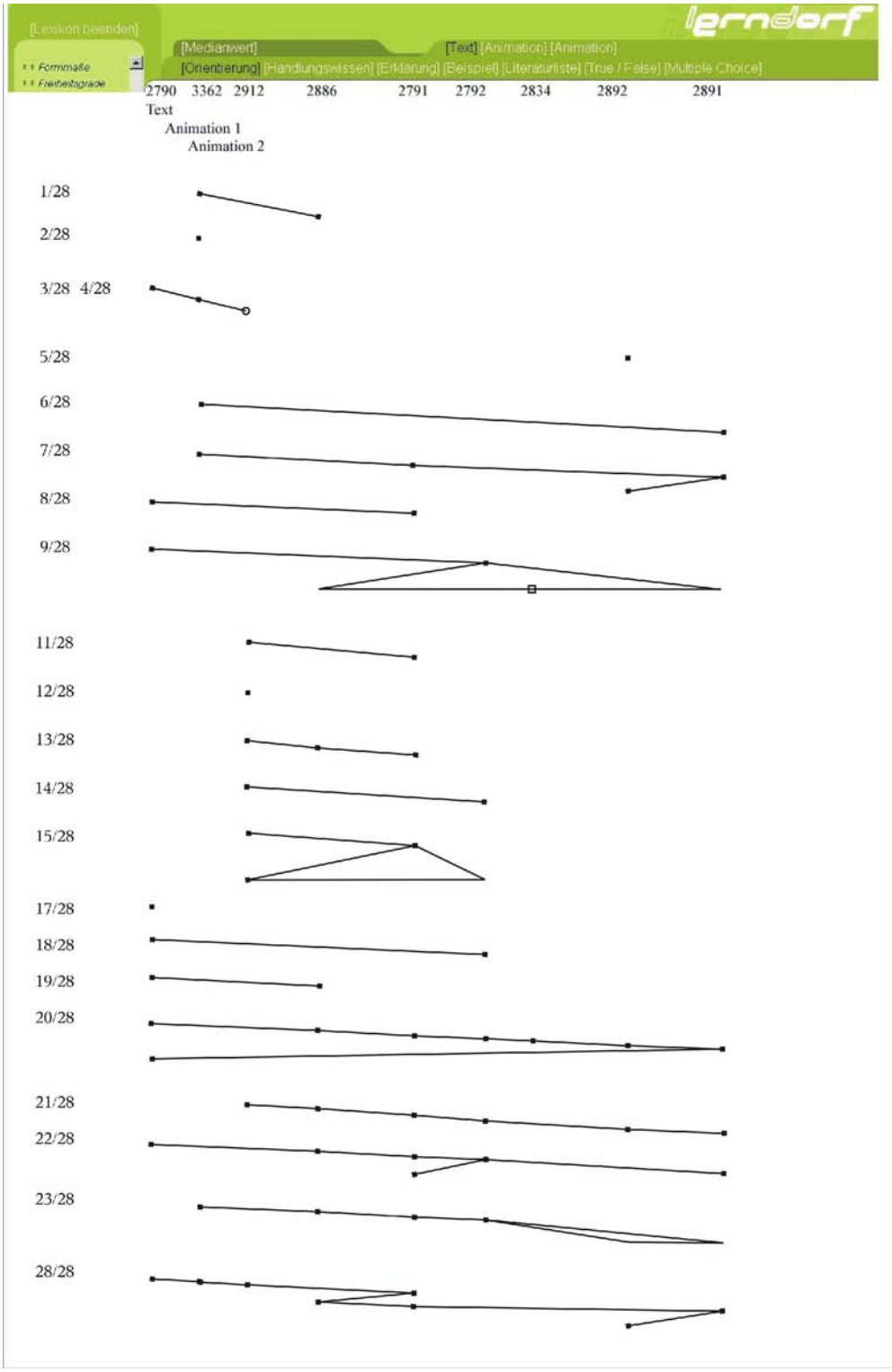

Abbildung 72: Navigationsmuster "Median" (516).

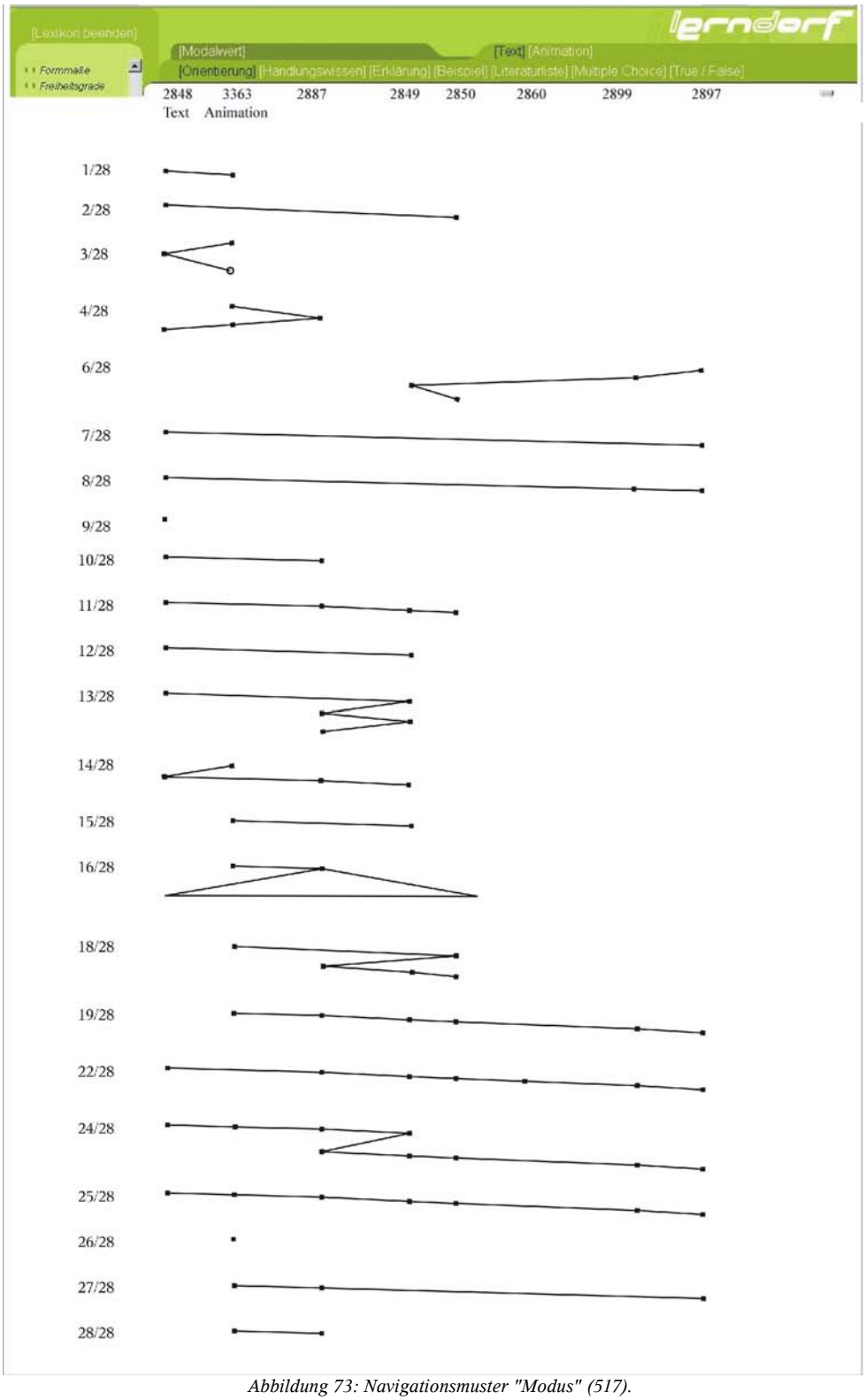

Abbildung 73: Navigationsmuster "Modus" (517).

11.2 Interpretationen der Navigationssequenzen

Aufbauend auf der formalen Darstellung der Clusterlösungen wird in diesem Abschnitt der Prozess der Mikronavigation auf *formaler* und *inhaltlicher* Ebene interpretiert: Welche Muster, Regelmäßigkeiten und Strukturen sind in den Navigationssequenzen enthalten? Welche inhaltlichen Aussagen können über die Navigationsmuster getroffen werden?

Grundlage der formalen Interpretation bilden die Navigationssequenzen als typische Abfolge von Wissenseinheiten. Grundlage der inhaltlichen Interpretation bilden die mit den Wissensarten verbundenen Metadaten. Dabei ist die inhaltliche Interpretation aufgrund der *Abstraktion* der Metadaten notwendigerweise *distanziert* von den konkreten Intentionen und Zielen der Nutzung (vgl. Kap. 2). Der typische Ablauf der Sequenz dient dabei als *Index* für zugrunde liegende Intentionen und Ziele. Die Navigationssequenzen werden dabei als konkrete Vorgehensweisen der Abbildung interpretiert, als *autodidaktisches Handeln* (vgl. Kap. 3.2, *Navigation als Autodidaktik*).

Auf formaler Ebene können die analysierten Navigationssequenzen zunächst hinsichtlich folgender grundlegender Merkmale gekennzeichnet werden:

- *Anzahl der Elemente der Sequenz*: Die Länge der Sequenz als Anzahl der Elemente beeinflusst die Levenshtein-Distanz und damit das Ergebnis der Clusteranalyse (vgl. Kap. 6). Ein grundlegendes Merkmal der Clusterlösung ist daher deren typische Sequenzlänge. Auf allgemeiner Ebene können Navigationssequenzen danach unterschieden werden, ob sie aus *einem*, *zwei, drei* oder *mehr* Elementen bestehen.
- *Lineare und nicht-lineare Navigation*: Bei der formalen Darstellung der Clusterlösungen wurde bereits auf die grafische Anordnung der Wissenseinheiten innerhalb der Lerneinheit hingewiesen. Mit Bezug auf diese grafische Darstellung kann eine *lineare* von einer *nicht-linearen* Navigation unterschieden werden: Die *lineare Navigation* folgt der Anordnung der Wissenseinheiten „von links nach rechts“ während eine *nicht-lineare Navigation* in einem Abweichen von dieser *links-rechts* Navigation besteht, wobei auf spezifische Wissenseinheiten *direkt* zugegriffen wird.
- *Fokussierung auf spezifische Wissenseinheiten*: Ähnlichkeit von Sequenzen besteht neben der Sequenzlänge vor allem in der Abfolge der Elemente (vgl. Kap. 5.1). Die Analyse des Navigationsprozesses fokussiert einerseits den wiederholten Aufruf von Wissenseinheiten und andererseits den Umstand, dass spezifische Wissenseinheiten gerade *nicht* in der Navigationssequenz enthalten sind.
- *Ausgangs- und Endpunkt der Sequenz*: In allgemeiner Perspektive können Navigationssequenzen nach dem Ausgangs- und nach dem Endpunkt der Sequenz unterschieden werden. Eine zentrale Bedeutung hat dabei das Orientierungswissen, da es bei dem Aufruf der Lerneinheit im Rahmen der

freien Navigation als erste Lerneinheit angezeigt wird (vgl. Orientierungswissen als *didaktische Station*, Kap. 3).[97]

- *Navigation im Rahmen eines Kurses (Kursnavigation)*: Bei der Darstellung der Konzeption der untersuchten Online-Lernumgebung wurde bereits darauf hingewiesen, dass zwei grundsätzliche Möglichkeiten der Navigation bestehen: die freie Exploration und das Folgen von Kursen. Die analysierten Sequenzen können daher bei der Navigationsanalyse danach unterschieden werden, ob sie im Rahmen freier Exploration oder der Kursnavigation entstanden sind.

11.2.1 Ein-Element-Sequenzen: „Überblick“

Analysiert man Navigationssequenzen hinsichtlich der Anzahl der Elemente, fallen zunächst diejenigen Cluster auf, die aus lediglich einem Element bestehen (*Ein-Element-Sequenzen*): Bei der Lerneinheit 513 sind dies vor allem die Cluster 1/28, 27/28; bei der Lerneinheit 515 die Cluster 1/28, 17/28; bei der Lerneinheit 516 die Cluster 2/28, 12/28, 17/28 und bei der Lerneinheit 517 die Cluster 9/28 und 26/28. Bezogen auf das Metadatum Wissensart handelt es sich bei diesen *Ein-Element-Sequenzen* um *Orientierungswissen* in unterschiedlichen Medientypen (Text, Animation). Diese Sequenzen bilden hinsichtlich der absoluten Zahlen und prozentualen Anteile die größten Cluster und sind in den jeweiligen Clusterlösungen als homogene Cluster enthalten.

Aus Sicht der Nutzenden stellt sich die *Ein-Element-Sequenz* folgendermaßen dar: Der Nutzer oder die Nutzerin hat die betreffende Lerneinheit aufgerufen, z.B. über den Themenbaum in der linken Navigationsleiste oder über eine Verknüpfung von einer anderen Wissenseinheit. Systemdefiniert wird die Wissenseinheit *Orientierung* als erste Wissenseinheit angezeigt. Die Nutzerin verlässt dann mit dem folgenden Klick diese Lerneinheit, d.h. es kommt zu keiner Navigationssequenz *innerhalb* der Lerneinheit. Aus inhaltlicher Perspektive sind vielfältige Interpretationen dieser Ein-Element-Sequenzen möglich, die im Folgenden angedeutet werden:

- Die Nutzerin verschafft sich einen Überblick über den Inhalt der Lernumgebung: Gibt es Lerneinheiten zu einem spezifischen Thema? Entsprechen die Inhalte der Lerneinheit den von der Nutzerin gesuchten Informationen? Als Ergebnis dieses Erkundungsprozesses entstehen dann *Ein-Element-Sequenzen* in der Lerneinheit.
- Die Nutzerin stellt im Rahmen der Erkundung anhand des Orientierungswissens fest, dass die Lerneinheit nicht die Information beinhaltet, die sie aktuell im Rahmen ihres Abbildungsprozesses be-

97 Bildet das Orientierungswissen nicht den Startpunkt der Navigationssequenz, kann dies auf unterschiedliche Gründe zurückzuführen sein, z.B. auf den Aufruf der Wissenseinheit im Rahmen der Kursnavigation, auf den Aufruf der Wissenseinheit durch die Option *Verlauf* des Browsers oder auf den Aufruf der Wissenseinheit über die Funktion *Suche*.

nötigt. Sie verlässt die Lerneinheit wieder und setzt seinen Navigationsprozess innerhalb der Lernumgebung fort, in dem sie eine andere Lerneinheit auswählt.

- Die Nutzerin stellt anhand des Orientierungswissens fest, dass die Lerneinheit zwar Informationen beinhaltet, die sie im Rahmen ihres Abbildungsprozesses benötigt, jedoch nicht zu diesem Zeitpunkt. Eine Navigation innerhalb der Lerneinheit zu einem späteren Zeitpunkt ist möglich (und im Rahmen der Navigationsanalyse auch analysierbar).
- Das Orientierungswissen der Lerneinheit wird als zentraler Ausgangspunkt von Navigationsprozessen genutzt. So befinden sich beispielsweise beim Orientierungswissen der Lerneinheit „Maße der zentralen Tendenz“ Verknüpfungen zu weiteren inhaltlich verwandten Lerneinheiten („Arithmetisches Mittel“, „Modus“, „Median“, u.a.). Die Nutzerin verwendet das Orientierungswissen also im Rahmen einer Speiche-Nabe-Navigation als Einstiegspunkt in das Thema „Maße der zentralen Tendenz“.

Gemeinsames Kennzeichen dieser inhaltlichen Interpretation ist die Verortung der *Ein-Element-Sequenzen* in Prozessen der *Makronavigation*, d.h. der Navigation zwischen Lerneinheiten. Im Zusammenhang der Darstellung der Sequenzanalyse und des Begriffes der Sequenz (vgl. Kap. 5) wurde bereits darauf hingewiesen, dass aus Perspektive der *Mikronavigation* dieser einmalige Aufruf keine Sequenz als Abfolge von Elementen darstellt.

11.2.2 Zwei-Elemente-Sequenzen: „gezieltes Nachschlagen“

Sequenzen als Abfolge von zwei Wissenseinheiten (*Zwei-Elemente-Sequenzen*) treten aus der Kombination von Orientierungswissen und einer weiteren Wissenseinheit oder dem Umschalten des Medientyps beim Orientierungswissen auf.

Geht man von den hier analysierten Lerneinheiten aus, liegen Kombination des Orientierungswissens mit folgenden weiteren Wissenseinheiten vor:

- *Handlungswissen* (*513*: 2/28, 21/28; *515*: 2/28, 14/28; *516*: 1/28, 19/28; *517*: 10/28, 28/28);
- *Erklärung* (*513*: 26/28; *515*: 1/28, 11/28; *516*: 11/28; *517*: 12/28; 15/28);
- *Beispiel* (*515:* 10/28; *516*: 14/28, 18/28; *517*: 2/28);
- *Literatur / Quellenwissen* (*513*: 2/28; *515*: 1/28; *516*: 2/28);
- *Multiple Choice* (*516*: 6/28);
- *True / False* (*517*: 7/28).

Die Kombination *Orientierungswissen – Aufgabe* bzw. *Entdeckende Aufgabe* ist im Rahmen der Zwei-Elemente-Sequenz nicht als eigenständige Clusterlösung enthalten.[98] In Hinblick auf die absoluten Zahlen und die prozentualen Anteile der Cluster stellt die Abfolge *Orientierungswissen – Handlungswissen* in Bezug auf die *Zwei-Elemente-Sequenzen* die häufigste eigenständige Clusterlösung dar.

Ein *Umschalten des Medientyps* bei der Wissensart Orientierung ist beispielsweise in folgenden Clusterlösungen als Zwei-Elemente-Navigation enthalten:[99]

- Text – Animation (*513*: 22/28; *515*: 5/28; *516*: 3/28; *517*: 1/28);
- Animation – Text (*515*: 3/28; *517*: 3/28).

Aus Sicht des Nutzers stellen sich *Zwei-Elemente-Sequenzen* folgendermaßen dar: Nach der Anzeige des Orientierungswissens (vgl. *Ein-Element-Sequenz*) wird der Navigationsprozess innerhalb der Lerneinheit fortgesetzt, d.h. es wird genau *eine* weitere Wissenseinheit ausgewählt. Die *Zwei-Elemente-Sequenzen* bestehen daher in der Regel aus dem Startpunkt Orientierungswissen (Text oder Animation) und einer weiteren Wissenseinheit. Diese weitere Wissenseinheit kann (wie oben formal beschrieben) aus einem direkten bzw. gezielten Zugriff auf eine der Wissenseinheiten der Lerneinheit bestehen.[100] Das zentrale Kennzeichen der *Zwei-Elemente-Sequenzen* besteht jedoch darin, dass nach dem Aufruf einer zweiten Wissenseinheit die Lerneinheit verlassen wird. Aus inhaltlicher Perspektive sind auch hier vielfältige Interpretationen möglich, die im Folgenden skizziert werden:

- Im Gegensatz zur *Ein-Element-Sequenz* stellt der Nutzer anhand des Orientierungswissens fest, dass die Lerneinheit Informationen beinhaltet, die er aktuell benötigt. Aus den zur Verfügung stehenden Wissenseinheiten wählt er dann genau eine weitere Wissenseinheit aus.
- Die zweite gewählte Wissenseinheit enthält die benötigten Informationen und der Nutzer verlässt die Lerneinheit.
- Die zweite gewählte Wissenseinheit enthält die benötigten Informationen nicht und der Nutzer geht davon aus, dass auch die weiteren in der Lerneinheit enthaltenen Wissenseinheiten diese Information nicht enthalten und verlässt die Lerneinheit.

Generell kann diese direkte bzw. zielgerichtete Auswahl einer Wissenseinheit als „gezieltes Nachschlagen" interpretiert werden: Ausgehend vom Orientierungswissen wählt der Nutzer genau eine spezifische Wissenseinheit und verlässt danach die Lerneinheit.

98 Diese Kombinationen ist aufgrund der in den analysierten Lerneinheiten enthaltenen Wissensarten lediglich in der Lerneinheit „Maße der zentralen Tendenz" (513) möglich.

99 Neben diesem Umschalten des Medientyps als *Zwei-Elemente-Navigation* ist auch ein mehrfaches Umschalten in den Clusterlösungen enthalten: Animation – Text – Animation (517: 3/28); Text – Animation – Animation (516: 4/28).

100 Im Fall der Auswahl des Handlungswissens als weiterer Wissenseinheit ist jedoch formal nicht eindeutig zu entscheiden, ob es sich um eine Zweischritt-Navigation „von links nach rechts" handelt oder um eine direkte, gezielte Auswahl der Wissenseinheit.

11.2.3 Drei- und Mehr-Elemente-Sequenzen: „Erkundung und Auseinandersetzung“

Sequenzen als Abfolge von drei Wissenseinheiten (*Drei-Elemente-Sequenzen*) kommen in vielfältigen Kombinationen vor, die im Folgenden skizziert werden. Auf die *Zwei-Elemente-Sequenz* als Kombination des Orientierungswissens mit einer weiteren Wissenseinheit wurde im vorangehenden Abschnitt hingewiesen. Die *Drei-Elemente-Sequenz* enthält eine darüber hinausgehende Wissenseinheit: Diese dritte Wissenseinheit kann wieder das Orientierungswissen als Ausgangspunkt der Navigationssequenz sein (*Orientierung – Wissenseinheit – Orientierung).* Dies ist beispielsweise in den folgenden Clusterlösungen der Fall: *Orientierung – Entdeckende Aufgabe – Orientierung (513*: 4/28); *Orientierung – Handlungswissen – Orientierung* (513: 20/28). Im Gegensatz zur vorangehenden Sequenz ist der Endpunkt der Sequenz nicht mit dem Anfangspunkt identisch, sondern die Sequenz wird durch eine weitere Wissenseinheit fortgesetzt (*Orientierung – Wissenseinheit 1 – Wissenseinheit 2).* Dies ist beispielsweise bei folgenden Clusterlösungen der Fall: *Orientierung – Handlungswissen – Erklärung* (513: 24/28) (516: 13/28) (517: 14/28); *Orientierung – True / False – Multiple Choice (515: 6/28).* Gemeinsamkeit besteht jedoch in der Anzahl von genau drei Wissenseinheiten.

Komplexer wird die Beschreibung von Sequenzen, die aus einer größeren Anzahl von Wissenseinheiten bestehen. Ein auffälliges Muster in diesen langen Sequenzen ist die „von links nach rechts“-Navigation. Variationen dieses Navigationsmusters bestehen darin, ob nach der „von links nach rechts“-Navigation der Navigationsprozess fortgesetzt wird oder nicht, ob bei einer Fortsetzung die Sequenz am Ausgangspunkt endet oder nicht, ob bei der „von links nach rechts“-Navigation einzelne Wissenseinheiten übersprungen werden oder nicht. Die folgenden Liste gibt einen Überblick über diese Muster innerhalb der „von links nach rechts“-Navigation:

- *ohne* Überspringen und *ohne* fortgesetzter Navigation (*515*: 18/28, 26/28; *517*: 22/28);
- *ohne* Überspringen und *mit* fortgesetzter Navigation,
- *mit* Rückkehr zum Ausgangspunkt (*513*: 5/28; 8/28; *515*: 19/28; *516*: 20/28);
- *ohne* Rückkehr zum Ausgangspunkt (*513*: 9/28; 10/28);
- *mit* Überspringen und *ohne* fortgesetzte Navigation (*513*: 7/28 ; *515*: 20/28; *516*: 21/28; 25/28; *517*: 19/28, 25/28);
- *mit* Überspringen und *mit* fortgesetzter Navigation (*515*: 21/28);
- *mit* Rückkehr zum Ausgangspunkt (*515*: 27/28).

Über dieses Muster der „von links nach rechts“-Navigation und dessen Variationen hinaus, die sich an einer linearen Navigation orientieren, sind vielfältige Muster mit nicht-linearem, direktem Zugriff auf Wissenseinheiten in den Clusterlösungen enthalten. Die Leistungsfähigkeit der vorliegenden Navigationsanalyse wird vor allem in Form differenzierter Clusterlösungen für diese Navigationsmuster deutlich, beispielsweise für

den Fall des Sequenzbeginns mit der Abfolge *Orientierung – Handlung - Erklärung*. So finden sich die folgenden Sequenzmuster in differenzierten Clusterlösungen:

- *Orientierung – Handlung – Erklärung* als Drei-Elemente-Sequenz (513: 24/28; 516: 13/28; 517: 14/28)
- *Orientierung – Handlung – Erklärung* mit fortgesetzter Navigation in Form von
    - *Quellenwissen* und weiterer heterogener Abfolge von Wissenseinheiten bei einer insgesamt langen Sequenz von Elementen in den Clusterlösungen *513*: 9/28 und 513: 10/28;
    - *Beispiel* ohne fortgesetzte Navigation als Vier-Elemente-Navigation in der Clusterlösung *517*: 11/28;
    - *Beispiel* mit fortgesetzter Navigation, z.B. in Form von Handlungswissen oder Multiple Choice in der Clusterlösung *516*: 22/28 oder in Form von Aufgaben in der Clusterlösung *516*: 23/28;
    - *Aufgabe* mit fortgesetzter Navigation, z.B. in Form von Erklärung oder Handlung in der Clusterlösung 513: 13/28.

Über diese differenzierte Clusterlösung für die Abfolge unterschiedlicher Elemente der Sequenz hinaus ist auch eine Differenzierung in Abhängigkeit des Ausgangspunktes bei ansonsten gleicher Abfolge der Elemente der Sequenz festzustellen: Sequenzen werden dementsprechend unterschiedlichen Clustern zugeordnet (vgl. 515: 18/28 und 26/28).

Darüber hinaus liegen differenzierte Clusterlösungen auch für weitere Muster der Navigationssequenzen vor, die im Folgenden inhaltlich auf Grundlage der Metadaten interpretiert werden:

- Muster: *Orientierung – Erklärung + weitere Wissenseinheiten*.
  Dieses Muster ist in der Clusterlösung 517: 13/28 in der Abfolge *Orientierung – Erklärung – Handlung – Erklärung – Handlung* enthalten, sowie in der Clusterlösung 516: 15/28 mit dem Sequenzbeginn *Orientierung - Erklärung*.
  Mit Bezug auf die Reihenfolge und die Fokussierung der Wissenseinheiten der Sequenzen kann dieses Muster als „erklärungsfokussiert“ bezeichnet werden.
- Muster: *Orientierung – Beispiel + weitere Wissenseinheiten.*
  Dieses Muster ist in der Clusterlösung 517: 18/28 in der Abfolge *Orientierung – Beispiel – Handlung – Erklärung – Beispiel* enthalten, sowie in der Clusterlösung 516: 9/28 und 515: 28/28 mit dem Sequenzbeginn *Orientierung - Beispiel*.
  Mit Bezug auf die Reihenfolge und die Fokussierung der Wissenseinheiten der Sequenzen kann dieses Muster als „beispielfokussiert“ bezeichnet werden.
- Muster: *Orientierung – Aufgabe + weitere Wissenseinheiten.*
  Variationen dieses Musters bestehen in unterschiedlichen Aufgabenarten, in diesem Fall *Multiple*

Choice und *True / False*. Dieses Muster ist in der Clusterlösung 515: 6/28 in dem Sequenzbeginn *Orientierung – Multiple Choice – True / False* (+ weitere Wissenseinheiten) enthalten; in der Clusterlösung 517: 8/28 als Drei-Elemente-Sequenz *Orientierung – Multiple Choice – True / False* sowie in der Clusterlösung 513: 4/28 in der Abfolge *Orientierung – Entdeckende Aufgabe – Orientierung*.
Mit Bezug auf die Reihenfolge und die Fokussierung der Wissenseinheiten der Sequenzen kann dieses Muster als „aufgabenfokussiert" bezeichnet werden.

- Muster: *Orientierung – weitere Wissenseinheiten – Aufgabe.*
  Dieses Muster ist in der Clusterlösung 515: 21/28 enthalten, das Ende der Sequenz besteht aus der Abfolge *Multiple Choice – True / False - Multiple Choice – True / False,* sowie in der Clusterlösung 516: 7/28 in der Abfolge *Orientierung – Handlung – Multiple Choice – True/False* und in der Clusterlösung 517: 27/28 in der Abfolge *Orientierung – Handlung – True / False*.

  Mit Bezug auf die Reihenfolge und die Fokussierung der Wissenseinheiten kann dieses Muster als „testfokussiert" bezeichnet werden.[101]

Aus Sicht der Nutzenden stellen sich Sequenzen mit mehreren Elementen folgendermaßen dar: Ausgehend vom Orientierungswissen navigieren sie durch Auswahl mehrerer Wissenseinheiten innerhalb der Lerneinheit. Es ist daher davon auszugehen, dass allgemein ein Interesse vorhanden ist, sich durch eine fortgesetzte Navigation mit den Informationen (Wissenseinheiten) der Lerneinheit auseinanderzusetzen. Bei dieser fortgesetzten Navigation können aus inhaltlicher Perspektive unterschiedliche Muster identifiziert werden.

Auf formaler Ebene wurde bereits auf das Muster der *linearen Navigation* „von links nach rechts" und dessen Variationen hingewiesen. Aus inhaltlicher Perspektive können diese Muster generell als *Erkundung* und *Auseinandersetzung* interpretiert werden. Die Nutzenden erkunden zunächst die Inhalte (Wissenseinheiten) der Lerneinheit und verschaffen sich einen Überblick, um sich gegebenenfalls anschließend mit spezifischen Wissenseinheiten auseinander zu setzen (vgl. 515: 19/28).

Allgemein kann diese *lineare Navigation* als ein Navigationsmuster interpretiert werden, das sich am Nutzungshabitus des Buches orientiert und der für das Buch typischen Leserichtung von links nach rechts folgt. Diese Nutzungsmuster des Buches werden auf Nutzungsmuster des Bereichs Online-Lernumgebung angewandt. Dabei folgt die Nutzenden einer durch das Medium festgelegten Reihenfolge, im Fall der Lernumgebung der Reihenfolge der grafischen Anordnung der Wissenseinheiten. Es liegt die Vermutung nahe, dass diese Muster der linearen Navigation vor allem für solche Nutzenden dominant ist, die über wenig Erfahrung im Umgang mit hypertextuellen, modularisierten Lernumgebungen verfügen.

Das Auslassen bzw. Überspringen von Wissenseinheiten stellt den Übergang zu *direkten, nicht-linearen* Navigationsmustern der gezielten Auseinandersetzung dar. Diese nicht-linearen Navigationsmuster wurden

101 Bei der „testfokussierten" Sequenz wird zuerst eine bzw. mehrere Wissenseinheiten aufgerufen und danach die Aufgabe. Bei der „aufgabenfokussierten" Sequenz wird zuerst eine Aufgabe und danach eine bzw. mehrere Wissenseinheiten aufgerufen.

oben als differenzierte Clusterlösungen dargestellt und als „erklärungsfokussiert“, „beispielfokussiert“, „aufgabenfokussiert“ und „testfokussiert“ interpretiert.

11.2.4 Navigationsmuster durch Fokussierung spezifischer Wissensarten

Interpretiert man die Sequenzen in Hinblick auf die Fokussierung von Wissensarten und in Hinblick auf Wissensarten, die gerade *nicht* in der Sequenz enthalten sind, werden weitere Muster erkennbar. Deutlichstes Beispiel hierfür ist das *Fehlen des Quellenwissen* und das *Fehlen der Aufgaben* in Navigationssequenzen.

Navigationsmuster, die als gemeinsames Kennzeichen gerade *kein Quellenwissen* als Element enthalten, können dabei allgemein als fokussiert auf die Auseinandersetzung mit dem in der Lerneinheit aktuell vorhandenen *Informationen* interpretiert werden. Hinweise auf weiterführende Quellen treten dabei in den Hintergrund. Vor allem die linearen Navigationsmuster, bei denen einzelne Wissenseinheiten übersprungen werden, können als Übergang zu nicht-linearen, direkten Navigationsmustern interpretiert werden. Besonders ist auch hier auf das Auslassen der Wissenseinheit *Quellenwissen* hinzuweisen (vgl. z.B. die Clusterlösungen 513: 7/28; 515: 20/28; 516: 21/28; 517: 19/28). Dieses Auslassen bzw. Überspringen kann einerseits als Hinweis auf Prozesse der gezielten und direkten Navigation interpretiert werden und andererseits darauf, dass im Rahmen gezielter Navigationsweisen das Quellenwissen nicht aufgerufen wird, da es keinen im engeren Sinne *explizit inhaltlichen* Beitrag zum Thema der entsprechenden Lerneinheit enthält. Das Auslassen bzw. Überspringen des Quellenwissens kann also unter dieser Perspektive als absichtsvolles, zielgerichtetes Navigationsmuster interpretiert werden.

Navigationsmuster, die als gemeinsames Kennzeichen *keine Aufgaben* (Entdeckende Aufgabe, Multiple Choice, True / False) als Elemente enthalten, können inhaltlich auf allgemeiner Ebene als „informationsfokussiert“ interpretiert werden (vgl. z.B. 515: 12/28). Im Fokus steht dabei die inhaltliche Auseinandersetzung mit den Wissenseinheiten, eine Überprüfung des Lernprozesses durch das Bearbeiten von Aufgaben ist kein Bestandteil dieses Navigationsmusters. Damit handelt es sich um ein der „Aufgabenfokussierung“ und „Testfokussierung“ entgegengesetztes Navigationsmuster.

11.2.5 Kursnavigation

Bei der Beschreibung der formalen Ergebnisse der Clusterlösungen wurde bereits auf einzelne Clusterlösungen als Bestandteil einer übergeordneten Kursnavigation hingewiesen. Zur generellen Identifizierung der Kursnavigation in den empirischen Sequenzen dient einerseits die definierte Abfolge der Wissenseinheiten, andererseits der Hinweis zur Navigationsweise in den Metadaten des Logfiles (vgl. Kap. 10.3: 128).

Grundsätzlich kann auf diese Weise im Rahmen der Navigationsanalyse die Kursnavigation detailliert analysiert werden, die neben der freien Exploration der Lernumgebung eine spezifische Navigationsstrategie darstellt. Die Reihenfolge der Bearbeitung kann jedoch nicht als Abbildungsprozess der Nutzenden interpretiert werden, da die Anordnung der Wissenseinheiten ja gerade durch den Autor des Kurses definiert wird.

Aus Perspektive der Navigationsanalyse als Analyse von Abbildungsprozessen ist die Kursnavigation jedoch in den Fällen von besonderem Interesse, in denen die definierte Reihenfolge des Kurses zugunsten einer freien Navigation verlassen wird. Diese Prozesse des Verlassens der definierten Reihenfolge des Kurses können im Rahmen der Navigationsanalyse zum Forschungsgegenstand werden, sind jedoch nicht Gegenstand der vorliegenden Arbeit.

Allgemein ist jedoch mit Blick auf die Kursnavigation feststellbar, dass nur wenige Nutzer und Nutzerinnen den Kurs „Statistik – Maße der zentralen Tendenz“ vollständig und in der vom Autor definierten Reihenfolge bearbeiten (vgl. Kap.17.17; *Teilkurs: „Statistik - Maße der zentralen Tendenz“*), sondern das Verlassen dieser Reihenfolge die dominante Bearbeitungsweise darstellt.

12 Ausblick: Variation, Weiterführung, Anknüpfungspunkte

Aufbauend auf der Methodologie der Navigationsanalyse mittels Optimal-Matching sind vielfältige *methodische* und *inhaltliche* Variationen und Weiterführungen sowie Anknüpfungspunkte möglich, die im Folgenden skizziert werden.

12.1 Methodische Variationen

Methodische Variationen der Navigationsanalyse beziehen sich auf die konkreten Parameter zur Durchführung der Optimal-Matching Analyse sowie der Kombination der Optimal-Matching Analyse mit weiteren Analyseverfahren. Die methodischen Variationsmöglichkeiten hinsichtlich der der Wahl der Clustermethode und der Clusteranzahl wurde bereits in Kapitel 9.3, *Formale Beschreibung der Clusteralgorithmen* ausführlich diskutiert und werden daher an dieser Stelle nicht weiter ausgeführt.

In Abhängigkeit der Forschungsfrage können allgemein folgende Parameter variiert und angepasst werden: Definition der Ereignisse bzw. Zustände der Sequenzen, erfasster Zeitraum und verwendete Zeitachse, Definition der Substitutionskosten sowie Definition der Indelkosten. Diese grundlegenden Variationen in der Durchführung der Optimal-Matching Analyse wurden in der vorliegenden Arbeit in Kapitel 5, *Sequenzdatenanalyse* und Kapitel 6, *Optimal-Matching Analyse* diskutiert.

Insbesonders ist im Zusammenhang der Ergebnisse der Navigationsanalyse darauf hinzuweisen, dass die Clusterlösungen auf Grundlage der Levenshtein-Distanz zunächst mathematische Ähnlichkeiten darstellen. Diese mathematische Ähnlichkeit ist jedoch nicht gleichzusetzen mit einer inhaltlichen Ähnlichkeit. Im Zusammenhang der Darstellung der formalen Clusterlösungen und deren Beschreibung wurde bereits darauf hingewiesen, dass die beschriebenen Clusterlösungen zwar aus mathematischer Perspektive homogene (und signifikante) Clusterlösungen darstellen, aus inhaltlicher Perspektive jedoch unterschiedliche Navigationsmuster in den Clusterlösungen enthalten sein können (vgl. 513: 2/28, 27/28; 515: 1/28; 516: 2/28). Dies ist insbesondere bei kurzen Sequenzen der Fall, wenn der Unterschied der Sequenzen aus lediglich einer Wissenseinheit besteht. Aus Perspektive des Optimal-Matching Verfahrens stellt der Unterschied einer Wissenseinheit eine geringe Distanz dar, die lediglich einer Operation entspricht. Im Verlauf der Clusteranalyse werden diese Sequenzen dann zu einem Cluster fusioniert, da sie sich nur gering voneinander unterscheiden.

Diese mathematische Interpretation der Sequenzen muss daher durch eine inhaltliche Interpretation ergänzt werden, wobei die Optimal-Matching Analyse gerade die Voraussetzung für diese inhaltliche Interpretation

von Mustern, Regelmäßigkeiten und Strukturen ermöglicht. Für den Fall mathematisch homogener, jedoch inhaltlich heterogener Clusterlösungen kann im Anschluss an die Clusteranalyse die Clusterzuordnung durch Rekodierung optimiert werden. Dabei können einerseits spezifische Muster einem anderen Cluster zugeordnet werden oder als Fälle eines neuen Clusters definiert werden.

12.2 Inhaltliche Variationen

Neben diesen methodischen Variationen bestehen vielfältige *inhaltliche* Variationen und Weiterführungen der Navigationsanalyse, die am Beispiel der Ausweitung auf Prozesse der Makronavigation, auf die Berücksichtigung der Verweildauer in den Zuständen, auf die Standardisierung der Sequenzlänge, auf die Analyse von Referenzsequenzen und auf die Ausweitung auf zusätzliche Daten des Navigierenden skizziert werden.

12.2.1 Ausweitung auf Prozesse der Makronavigation

Eine inhaltliche Variation der Navigationsanalyse besteht in der Ausweitung der Analyseperspektive von Prozessen der Mikronavigation (als Navigationsprozesse *innerhalb* von Lerneinheiten) auf *Prozesse der Makronavigation* (als Navigationsprozesse *zwischen* Lerneinheiten). Dabei kann die *Makronavigation* in Abhängigkeit der Forschungsfrage auf unterschiedlichen Stufen der Abstraktion analysiert werden: als Sequenz von Lerneinheiten *ohne* Berücksichtigung der Navigation innerhalb dieser Wissenseinheiten oder aber als Sequenz von Lerneinheiten *mit* Berücksichtigung der Navigation innerhalb dieser Wissenseinheiten. Für diese inhaltliche Variation und Ausweitung der Navigationsanalyse ist eine veränderte Datenaufbereitung notwendig. Eine Veränderung der Form der Datenerhebung in der Lernumgebung ist für diese inhaltliche Variation nicht notwendig.

12.2.2 Ausweitung auf die Verweildauer in Zuständen

Weiter oben wurde bereits ausgeführt, dass in der vorliegenden Navigationsanalyse *Zeit* als zeitlicher Verlauf der Abfolge von Elementen berücksichtigt wird. Eine weitere Möglichkeit der Variation der Navigationsanalyse besteht in der Berücksichtigung von *Zeit* als Verweildauer in den einzelnen Zuständen (Wissenseinheiten).

12.2.3 Ausweitung der Standardisierung der Sequenzlänge

Eine weitere inhaltliche Variation betrifft die Länge der Sequenzen und die Frage der Standardisierung. So kann beispielsweise aus analytischen Gründen die Sequenzlänge auf die ersten 3 oder 4 Zustände (Wissenseinheiten) begrenzt werden. Ein Vergleich dieser zensierten Sequenzen mit dem Ergebnis der Clusteranalyse der vollständigen Sequenzen kann dann Auskunft darüber geben, ob ein Zusammenhang besteht, d.h. ob anhand der ersten 3 oder 4 Klicks eine Prognose der folgenden Navigationssequenz möglich ist.

12.2.4 Ausweitung auf die Analyse von Referenzsequenzen

Darüber hinaus kann der paarweise Sequenzvergleich der Optimal-Matching Analyse zur Errechnung der Levenshtein-Distanz variiert werden. TDA bietet die Möglichkeit der Verwendung von Referenzsequenzen, d.h. die Sequenzen des Datensatzes werden nicht paarweise sondern mit einer *Referenzsequenz* verglichen. In der TDA-Syntax wird dieses „Pattern Matching“ (Rohwer / Pötter 2005: 471) durch den *seqpm*-Befehl (*seq*uence *p*attern *m*atching) umgesetzt.

```
seqpm (
    sn=...,          number of sequence data structure, def. 1
    ps=...,          definition of patterns
    df=...,          test output file
    nfmt=...,        integer print format, def. 4
    v=...,           additional variables for output file
    dtda=...,        TDA description for output file
) = output_file_name;
```

Abbildung 74: Syntax des seqpm-Befehls (TDA-Manual 2005: 471).

Bei dieser *Referenzsequenz* kann es sich sowohl um eine *empirische*, d.h. um eine im aktuellen Sequenzdatensatz vorhandene Sequenz handeln als auch um eine *theoriebasierte* Sequenz.
Durch Verwendung einer *empirischen Referenzsequenz* kann beispielsweise das Ergebnis der Clusteranalyse validiert und optimiert werden. In Kapitel 9.2, *Clusteranalyse im Rahmen der Navigationsanalyse* wurde bereits ausgeführt, dass eine Validierung und Optimierung der Ergebnisse einer hierarchischen Clusteranalyse durch partitionierende Verfahren wie *k-Means* im Rahmen der Navigationsanalyse nicht möglich sind. Eine Optimierungsmöglichkeit eröffnet jedoch die Verwendung empirischer Referenzsequenzen. Dazu wird in einem ersten Schritt für jedes Cluster der hierarchischen Clusterlösung durch inhaltlich-formale Interpretation die clustertypische Sequenz bestimmt.

Diese clustertypische Sequenz bildet dann die Referenzsequenz: Alle Sequenzen des Datensatzes werden mit dieser Sequenz verglichen. Das Ergebnis besteht aus einer Liste, die für jede Sequenz die Distanz zur Referenzsequenz enthält. Zur Validierung der Clusterlösung wird dann dieses Ergebnis mit dem Ergebnis der Clusteranalyse verglichen. Auf der gleichen Grundlage kann das Ergebnis der Clusteranalyse auch optimiert werden, in dem Sequenzen aufgrund des Ergebnisses des Vergleichs mit der Referenzsequenz einer anderen Clusterlösung zugeordnet werden. Analog zum Vorgehen des *k-Mean*s Clusterfahrens kann auf diese Weise die Clusterlösung hierarchischer Clusterverfahren optimiert werden, deren grundlegendes Kennzeichen gerade die Nichtrevidierbarkeit der Zuordnung von Fällen zu Clustern ist (vgl. Kap. 9.2.2).
Durch die Verwendung *theoriebasierter Referenzsequenzen* kann die Distanz empirischer Sequenzen zu theoretisch entwickelten Sequenzen analysiert werden. Als Bezugspunkte der vorliegenden Navigationsanalyse können beispielsweise didaktische Modelle als idealtypische Sequenzen verwendet werden (vgl. Meder 2006). In Anlehnung an Weber (1990) können als *Idealtypen* gerade solche Sequenzen aufgefaßt werden, die nicht in reiner Form in der sozialen Wirklichkeit zu finden sind. Die Verwendung solcher *idealtypischen Referenzsequenzen* ist gerade aus diesem Grund aufschlussreich, da das Ergebnis des Vergleichs die Distanz zur dieser Referenzsequenz ausdrückt und eine vollständige Übereinstimmung nicht notwendig ist.

12.2.5 Ausweitung auf zusätzliche Daten des Navigierenden

Eine Weiterführung der Navigationsanalyse besteht in der Erhebung und Analyse von Daten, die über die reinen Navigationssequenzen hinausgehen und sich beispielsweise auf das Resultat, das Ziel oder die Intention des Navigationsprozesses bzw. auf soziodemografische Daten der Nutzenden beziehen. Ganz allgemein kann auf der Grundlage weiterer Daten (Variablen) die Clusterzugehörigkeit der Sequenzen in spezifischen Forschungsdesigns als *abhängige* und *unabhängige* Variable für weitergehende Analysen verwendet werden (vgl. Kap. 4.2; *Web-Mining)*. Durch die Kombination der Navigationsanalyse insbesondere mit multinominalen logistischen Modellen können aufbauend auf diesen Daten z.B. Zusammenhänge zwischen Nutzertyp (z.B. Geschlecht oder Bildung) und Navigationsstrategien analysiert werden.

12.3 Triangulation

Neben den beschriebenen methodologischen und inhaltlichen Variationsmöglichkeiten und der Ausweitungen der Navigationsanalyse ist auf deren Einbettung im Rahmen der Triangulation von Text-, Ton- und Bilddaten hinzuweisen (vgl. Kap. 2.2: 11; *Triangulation von Text-, Ton- und Bilddaten*). Vor allem die

Kombination der Navigationsanalyse mit weiteren statistischen Verfahren ermöglicht dabei eine vertiefte Analyse von Navigationsprozessen.

Durch die Kombination der Navigationsanalyse mit Verfahren der *Ereignisdatenanalyse* (vgl. Kapitel 8) kann der explorativ-heuristische Ansatz der Navigationsanalyse durch das Potential der Analyse bedingender Faktoren (Kausalanalyse) auf Grundlage der Berechnung von Übergangswahrscheinlichkeiten ergänzt und erweitert werden. Damit wird die Navigationsanalyse ausgeweitet auf die Überprüfung von theoretischen entwickelten Modellen und Hypothesen.

Darüber hinaus stellt der Bereich des *Data-Mining* ein breites methodisches Repertoire zur Verfügung (vgl. 4.2, Web-Mining), dass bisher bei der Analyse von Navigationsprozessen im Bereich von Online-Lernumgebungen noch wenig Berücksichtigung findet.

12.4 Anknüpfungspunkte

In diesem Abschnitt werden Anknüpfungspunkte der Navigationsanalyse mittels Optimal-Matching dargestellt, um das analytische Potenzial für den Bereich des E-Learning beispielhaft zu skizzieren. Dabei wird vor allem Bezug genommen auf die Diskussionen um *Learning Design Patterns*, auf *Social Software* und *Social Navigation*, sowie auf die Diskussion von *Qualität im E-Learning aus Lernerperspektive* und auf die Entwicklung differenzierter, pädagogisch-didaktischer Empfehlungssysteme.

Im Kontext der Standardisierungsbestrebungen im E-Learning und vor allem im Kontext des IMS Global Learning Consortiums gewinnt die Identifizierung von *Learning Design Patterns* bzw. *Pedagogical Patterns* zur Konzeption von E-Learning Angeboten zunehmend an Bedeutung und stellt nach Koper (2006) eine zentrale Herausforderung an die Forschung dar: „The idea of learning design patterns and the possibility to recognise them automatically with pattern detection algorithms is a new field of work that is worthwile to elaborate in future“ (Koper 2006). Generell wird zur Identifizierung pädagogischer Muster ein *deduktives* von einem *induktivem* Vorgehen unterschieden (vgl. Brouns u.a. 2005). Während beim deduktiven Vorgehen Muster von Experten entwickelt werden (vgl. Instructional Design; Reigeluth 1983, 1999), basiert das induktive Vorgehen auf der Analyse der didaktischen Struktur bestehender (Online-)Kurse, um pädagogische Muster zu identifizieren und zu extrahieren. Das Potenzial der Sequenzanalyse mittels Optimal-Matching ist vor allem auch im Bereich dieses induktiven Vorgehens zu verorten. Dabei kann der induktive Ansatz auf die Analyse von Muster empirischer Navigationsprozesse in hypermedialen Online-Lernumgebungen ausgeweitet werden.

Einen weiteren Anknüpfungspunkt stellt der Ansatz der *Social Navigation* dar (vgl. Dourish / Chalmers 1994; Höök / Benyon / Munro 2003). Den Ausgangspunkt und das Ziel dieser Forschungsperspektive fasst

Dieberger (2003: 293-294) folgendermaßen zusammen: „Although many people may access an information system at the same time, most systems maintain the illusion of a dedicated resource and the only indication of a large number of users simultaneously accessing a system might be an unusually slow response time. [...] A goal of social navigation is to utilise information about other people's behaviour for our own navigational decisions“. Dieses Nutzen von Informationen über den Navigationsprozess anderer Nutzender kann dabei direkt oder auch indirekt stattfinden. *Direkte soziale Navigation* setzt die Kopräsenz der Nutzenden in der Lernumgebung voraus und besteht in einem direkten Kontakt, wohingegen indirekte soziale Navigation auf Informationen über die Interaktion anderer Nutzender mit der Lernumgebung beruht (z.B. besonders effektive oder attraktive Wege, viel- oder wenig genutzte Pfade). Für die Konzeption dieser *indirekten sozialen Navigation* stellt die Sequenzanalyse mittels Optimal-Matching ein besonderes Potential dar, weil Navigationssequenzen als Verläufe in die Konzeption der indirekten sozialen Navigation einbezogen werden können.

Die Kennzeichnung der Hypertext-Technologie als Pull-Medium mit der Notwendigkeit des Entfaltens ist Ausgangspunkt eines relationalen Qualitätsverständnisses des E-Learning und einer lernerbezogenen Perspektive (vgl. Ehlers 2002; Ehlers 2004). Grundlegende These ist, dass einem Lernmedium nicht bereits vorab eine Lernqualität als solche zugeschrieben werden kann, sondern diese erst im Prozess des Lernens entsteht und wesentlich vom Lerner mitbestimmt wird (vgl. Ehlers 2004). Dem Lernenden kommt bei diesem Prozess der Konstitution von Qualität die Rolle des Koproduzenten zu: Die Online-Lernumgebung liefert beispielsweise den Inhalt und Kommunikationswerkzeuge, der Lernende muss jedoch selbst tätig werden. So betont Ehlers (2002: 9) die zentrale Bedeutung dieser Interaktion für die Konstitution von Qualität: „Qualität entsteht erst dann, wenn der Lernende mit dem Lernarrangement in Interaktion tritt. Erst dann, wenn gelernt wird entsteht auch Lernqualität (Ko-Produktion des Lernerfolges). Ein E-Learning-Lernarrangement hat keine Lernqualität an sich. Es ist lediglich der Rahmen (das Arrangement) mit Hilfe dessen sich der Lernprozess vollzieht.“ Dabei unterscheidet Ehlers verschiedene Qualitätsebenen: die Voraussetzungen (‚Inputqualität’), den Lernprozess (‚Prozessqualität’) und das Ergebnis (‚Outcomequalität’). Insbesondere die in dieser Arbeit dargestellte Analyse von Navigationsprozessen bezieht sich auf die Relation von Lernendem, Lernumgebung und Lerninhalt. Eine konsequente Qualitätsforschung des E-Learning unter Berücksichtigung der Lernerperspektive muss also neben der ‚Inputqualität’ und der ‚Outcomequalität’ vor allem die beschriebenen Relationen der ‚Prozessqualität' berücksichtigen.

Die Methodologie der Analyse von Navigationsprozessen mittels Optimal-Matching leistet einen zentralen Beitrag für die Qualitätsentwicklung von E-Learning unter Prozessperspektive (‚Prozessqualität’), d.h. hinsichtlich der Relation von Lernendem, Lernumgebung und Lerninhalt. Die Lernenden und ihre Handlungen werden zum zentralen Gegenstand. Damit kommen Fragen der Interaktivität in den Fokus und besonders Fragen des Potenzials von Prozessen der Rückkopplung und des Feedback zur Unterstützung von Online-Lernprozessen. Die Kenntnis der konkreten Navigationsprozesse ist dabei die Voraussetzung für differenzierte Rückmeldungen – personal sowie digital – und bildet die Grundlage für Strategien der Mikro-Adapta-

tion (vgl. Leutner 1992). Gerade neuere Entwicklungen wie Web 2.0, Social Software und Social Navigation versprechen neuartige und vielfältige Möglichkeiten der Art, des Umfangs und des Zeitpunktes einer lernförderlichen Rückkopplung. So tritt bei dem Konzept *Social Software* der kooperative Aspekt des E-Learning in den Vordergrund und geht damit weit über die 1 : 1 Situation eines isolierten Lerners vor einem Computer hinaus: Lernen wird zunehmend als sozialer Prozess verstanden, als Lernen *in* einer Gruppe und Lernen *von* einer Gruppe.

Insbesondere stellt die Kenntnis konkreter Navigationsprozesse und deren Analyse die Voraussetzung für die Entwicklung differenzierter pädagogisch-didaktischer Empfehlungssysteme als spezifische Form lernförderlicher Rückkopplung dar. Analog zu *amazon.de* interpretiert ein Empfehlungssystem die Handlungen der Nutzenden und gibt auf Grundlage dieser Interpretation Empfehlungen, die für den einzelnen Nutzer oder die einzelne Nutzerin hilfreich sind, d.h. sie in ihren E-Learningprozessen hilfreich unterstützen. Gegenwärtige Empfehlungssysteme auf der Grundlage von Assoziationsanalysen sind vor allem aus dem Bereich des E-Commerce bekannt, z.B. als *Warenkorbanalyse*. Dabei steht die Analyse von Nutzungsinformationen im Hinblick auf gemeinsame Interesse der Nutzer im Vordergrund. Auf Grundlage dieser *Warenkorbanalyse* werden Kaufempfehlungen abgeleitet, wie dies z.B. bei *amazon.de* als dynamisches Empfehlungssystem implementiert ist: „Kunden, die diesen Artikel gekauft haben, kauften auch:“, „Kunden, die diesen Artikel angesehen haben, haben auch angesehen:“, „Unser Vorschlag: Kaufen Sie jetzt diesen Artikel zusammen mit“. Allerdings unterscheiden sich die Aktivitäten eines Käufers von denen eines Lernenden, der Kaufprozess unterscheidet sich vom Lernprozess, das Ergebnis eines Lernprozesses unterscheidet sich vom Ergebnis eines Kaufprozesses. Darüber hinaus bleibt z.B. die Frage offen, ob ein pädagogisch-didaktisches Empfehlungssystem ähnliche Informationen vorschlägt - wie dies z.B. bei *amazon.de* der Fall ist - oder aber im Sinne einer absichtsvollen Irritation abweichende bzw. konträre Informationen. Grundlage eines solchen pädagogischen Empfehlungssystems ist jedoch in jedem Fall die Analyse und Interpretation der Handlungen der Nutzenden, der E-Learningprozesse.

Die Berücksichtigung von *Prozesse*n des E-Learning ist insbesondere dann notwendig, wenn mit Online-Lernumgebungen die Vermittlung prozeduralen Wissens und tätigkeitsorientierter Kompetenzen angestrebt wird. Dabei zielt die Analyse von E-Learningprozessen mittels Optimal-Matching auf den Kern einer Didaktik als Handlungswissenschaft, die die konkret-empirische Abbildung von Raumgestalten in Zeitgestalten (Didaktik) bzw. die Abbildung von Zeitgestalten in Raumgestalten (Autodidaktik) analysiert und hinsichtlich Adäquatheit und alternativer Möglichkeiten reflektiert (vgl. Meder 2003; Meder 2006). Gerade Hypertext als grundlegende Technologie von Online-Lernumgebungen stellt einen radikalen medialen Strukturwandel dar, in dem bisherige Prozesse der Abbildung grundlegend zur Disposition stehen.

Grundlage der beschriebenen Ansätze ist jedoch in jedem Fall die Analyse und Interpretation der Navigation von Nutzenden in online-Umgebungen, d.h. die Berücksichtigung der konkreten Prozesse der Navigation.

13 Zusammenfassung und Ausblick

Den Ausgangspunkt der vorliegenden Arbeit bildete die Frage nach der methodologischen Grundlage, auf der Navigationsprozesse in Online-Lernumgebungen analysiert werden können.

Die im Rahmen dieser Arbeit entwickelte explorativ-heuristische Navigationsanalyse beantwortet diese Frage mit Verweis auf eine Methodologie, die Navigationsprozesse als Sequenzen analysiert. Ziel ist das Identifizieren und Gruppieren von Muster, Regelmäßigkeiten und Strukturen in Navigationssequenzen. Dabei wird der Prozess der Auseinandersetzung von Nutzern in Online-Lernumgebungen bis auf die Ebene einzelner Mausklicks analysierbar. Das analytische Potenzial der Methodologie wurde anhand der Analyse von Navigationssequenzen von Nutzenden einer hypertextuellen, metadatenbasierten Online-Lernumgebung demonstriert.

Die Methodologie der Navigationsanalyse wurde in einem triangulativen Ansatz auf der Grundlage von Ton-, Bild- und Textdaten verortet. Die Analyse von Navigationssequenzen fokussiert in diesem Rahmen den Bereich der Analyse von Sequenzdaten als Textdaten. Diese Textdaten beruhen auf den *sequenzierten Logdaten* der Handlungen von Nutzenden in Online-Lernumgebungen: In den Textdaten wird der Navigationsprozess einzelner Nutzender in der Lernumgebung als Sequenz abgebildet.

Den methodischen Kern der Navigationsanalyse bildet das Optimal-Matching Verfahren in Verbindung mit Verfahren der Clusteranalyse: Anhand der grundlegenden Operationen *Einfügen („insertion“)*, *Löschen* („deletion“) und *Austauschen* („substitution“) wird die Levenshtein-Distanz für den paarweisen Vergleich aller Sequenzen bestimmt. In der Levenshtein-Distanz als Maßzahl kommt die Distanz der verglichenen Sequenzen zum Ausdruck. Die Levenshtein-Distanz dient als Ausgangspunkt clusteranalytischer Verfahren, um die analysierten Sequenzen in homogene Gruppen zu fusionieren.

Im Rahmen der Entwicklung und Begründung der Methodologie der Navigationsanalyse wurden grundlegende Parameter überprüft: Für das Verfahren der Optimal-Matching Analyse wurde der Effekt unterschiedlicher Definitionen der Substitutionskosten empirisch analysiert und die verwendete Definition der Substitutionskosten inhaltlich und formal begründet. Dabei wurde der Effekt der Definition der Substitutionskosten auf das konkrete Vorgehen des Optimal-Matching Algorithmus sowie auf den konkreten Wert der Levenshtein-Distanz analysiert.

Für das Verfahren der Clusteranalyse wurde der Effekt unterschiedlicher Clusteralgorithmen empirisch analysiert und die Wahl des verwendeten Clusteralgorithmus inhaltlich und formal begründet. Darüber hinaus wurde die im Rahmen der Navigationsanalyse verwendete Anzahl der Cluster begründet.

Die Ergebnisse der Navigationsanalyse machen deutlich, dass auf dieser methodologischen Grundlage die differenzierte Analyse von Prozessen der Navigation in Online-Lernumgebungen möglich ist. Die Clusterlösungen sind sowohl in formaler als auch in inhaltlich-didaktischer Hinsicht interpretierbar und verweisen auf in den Navigationssequenzen enthaltene Muster, Regelmäßigkeiten und Strukturen.

So ist beispielsweise die Identifizierung von linearen Navigationsmustern möglich, sowie von Mustern der „Erkundung“ und der „Auseinandersetzung“. Darüber hinaus sind nicht-lineare Muster als direkte und gezielte Navigation identifizierbar. Diese wurden auf Grundlage einer formalen Beschreibung sowie der Metadaten als „erklärungsfokussierte“, „beispielfokussierte“, „aufgabenfokussierte“ und „testfokussierte“ Navigationsmuster interpretiert und den Strategien „Überblick“ und „Nachschlagen“ zugeordnet. Neben den linearen und nicht-linearen Navigationsmustern wird auch die Fokussierung des Navigationsprozesses auf spezifische Wissenseinheiten analysierbar, z.B. in Form von Sequenzmustern, in denen spezifische Wissenseinheiten gerade *nicht* enthalten sind.

Über generelle Muster, Regelmäßigkeiten und Strukturen hinaus wird durch die Ergebnisse der Navigationsanalyse die Vielfalt konkreter empirischer Navigationssequenzen deutlich. Diese generelle Heterogenität empirischer Navigationssequenzen kann als Hinweis auf die Qualität der analysierten Lernumgebung interpretiert werden. Die Lernumgebung ist konzeptionell und praktisch in der Lage, eine Vielzahl unterschiedlicher Navigationssequenzen und unterschiedlicher Navigationsmuster abzubilden.

Allgemein können die identifizierten Muster der *Mikronavigation* als Elemente einer übergeordneten *Makronavigation* interpretiert werden. Die Muster der Mikronavigation bilden dabei die *Bausteine* für Prozesse der Makronavigation. Auf der Ebene der Mikronavigation wird die Aneignung eines einzelnen Themas fokussiert, auf der Ebene der Makronavigation die Aneignung eines umfassenderen Gegenstandsbereichs. Somit können insbesondere Strategien und Metaregeln der Navigation in hypertextuellen Lernumgebungen analysiert werden. Auf dieser Grundlage können dann beispielsweise theoretische Modelle der Navigation durch ein induktives Vorgehen entwickelt werden. Auf Grundlage der in dieser Arbeit identifizierten komplexen Muster der Mikronavigation sowie der konkreten Navigationssequenzen wird erkennbar, welchen hohen Komplexitätsgrad darauf aufbauende Prozesse der Makronavigation grundsätzlich erreichen können.

Die Methodologie der Navigationsanalyse in Online-Lernumgebungen ermöglicht die differenzierte Analyse von Prozessen der Interaktion von Nutzern mit Online-Lernumgebungen, die in der gegenwärtigen Forschung zu E-Learning in dieser Weise bislang keine Berücksichtigung finden. Die Analyse *sequenzierter* Verlaufsdaten im Rahmen der Navigationsanalyse setzt genau an dem Punkt an, an dem die Analyse *aggregierter* Logdaten endet: Der Navigationsverlauf als Sequenz wird zum Ausgangspunkt und zur Analyseeinheit. Generell kann auf der Grundlage der Navigationsanalyse eine sehr große Anzahl von Sequenzen miteinander verglichen werden, was die Sequenzlänge wie auch die Anzahl unterschiedlicher Elemente betrifft.

Während in der Diskussion um E-Learning vor allem die Aspekte der räumlichen und zeitlichen Unabhängigkeit des Lernens im Vordergrund stehen, fokussiert die Navigationsanalyse die konkreten Lern*wege*

der Nutzenden und interpretiert diese Prozesse der Navigation als das *lineare Entfalten eines nicht-linearen Hypertextes*. Dieses lineare Entfalten wird aus pädagogisch-didaktischer Sicht als *autodidaktisches Handeln* interpretiert. Analysiert wird dabei analog zum *zweiten Hauptsatz der Didaktik* die empirische Abbildung des zeitlich verlaufenden Lernprozesses in den sachlogischen Raum der Bedeutungsbeziehungen. Gerade Hypertext als grundlegende Technologie von Online-Lernumgebungen stellt einen radikalen medialen Strukturwandel dar, in dem bisherige Prozesse der Abbildung grundlegend zur Disposition stehen. Auf Grundlage der Methodologie der Navigationsanalyse wird dieses *autodidaktische Handeln* rekonstruierbar und analysierbar. Dabei zielt die Analyse von E-Learningprozessen mittels Optimal-Matching auf den Kern einer Didaktik als Handlungswissenschaft, die die konkret-empirische Abbildung von Raumgestalten in Zeitgestalten (Didaktik) bzw. die Abbildung von Zeitgestalten in Raumgestalten (Autodidaktik) analysiert und hinsichtlich Adäquatheit und alternativer Möglichkeiten reflektiert. Damit geht die Navigationsanalyse als Sequenzanalyse auf methodischer Ebene im Bereich des E-Learning weit über die gegenwärtige Berücksichtigung aggregierter Logdaten hinaus.

Die Fokussierung der Navigationsprozesse *während* des E-Learning ist von besonderer Bedeutung, da die Kenntnis von Navigationsprozessen als autodidaktischem Handeln vielfältige Anknüpfungspunkte didaktischen Handelns ermöglicht und darüber hinaus gerade Bildungs- und Lern*prozesse* den grundlegenden Gegenstandsbereich der Pädagogik bilden. Kenntnisse über den Ablauf von Prozessen des E-Learning lassen vielfältige Anknüpfungspunkte für pädagogisches Handeln erwarten, was die didaktische Konzeption von Online-Lernumgebungen wie auch die Unterstützung von Lernenden betrifft. Die Methodologie der vorliegenden explorativ-heuristischen Navigationsanalyse mittels Optimal-Matching bildet somit auf methodologischer wie forschungspraktischer Ebene eine Grundlage der Analyse vielfältiger erziehungswissenschaftlicher Fragestellungen im Bereich des E-Learning.

14 Literaturverzeichnis

Abbott, Andrew; Forrest, John (1986): "Optimal Matching Methods for Historical Sequences". Journal of Interdisciplinary History, 16, 3. S. 471-494.

Abbott, Andrew; Hrycak, Alexandra (1990): "Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers". The American Journal of Sociology, 96, 1. S. 144-185.

Abbott, Andrew (1990a): "A Primer on Sequence Methods". Organization Science, 1, 4. S. 375-392.

Abbott, Andrew (1990b): "Conception of time and events in social science methods". Historical studies, 23, 4.

Abbott, Andrew (1992): "From Causes to Events: Notes on Narrative Positivism". Sociological Methods & Research, 20, S. 428-455.

Abbott, Andrew (1995a): "Sequence Analysis: New Methods for Old Ideas". Annual Review of Sociology, 21, S. 93-113.

Abbott, Andrew (1995b): "A comment on 'Measuring the Agreement between Sequences'". Sociological Methods & Research, 24, 2. S. 232-243.

Abbott, Andrew; Barman, Emily (1997): "Sequence Comparison via Alignment and Gibbs Sampling: A formal analysis of the emergence of the modern sociological article". Sociological Methodology, 27, S. 47-87.

Abbott, Andrew; Tsay, Angela (2000a): "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect". Sociological Methods & Research, 29, 1. S. 3-33.

Abbott, Andrew (2000b): "Reply to Levine and Wu". Sociological Methods & Research, 9, 1. S. 65-87.

Aisenbrey, Silke (2000): Optimal Matching Analyse: Anwendungen in den Sozialwissenschaften. Leske und Budrich: Opladen.

Backhaus, Klaus; Erichson, Bernd; Plinke, Wulff; Weiber, Rolf (2000): Multivariate Analysemethoden: eine anwendungsorientierte Einführung. Springer: Berlin.

Baur, Nina (2005): Verlaufsmusteranalyse: methodologische Konsequenzen der Zeitlichkeit sozialen Handelns. VS, Verlag fur Sozialwissenschaften: Wiesbaden.

Berendt, Bettina; Brenstein, Elke (2001): Visualizing Individual Differences in Web Navigation. <http://www.wiwi.hu-berlin.de/~berendt/berendt_brenstein_BRMIC.pdf>, (28.08.2006).

Berendt, Bettina; Spiliopoulou, M. (2002): Assoziations- und Pfadanalyse. <http://www.wiwi.hu-berlin.de/~berendt/Papers/berendt_spiliopoulou_02_theorie.pdf>, (28.08.2006).

Bergmann, Jörg (1985). "Flüchtigkeit und methodische Fixierung sozialer Wirklichkeit." In: W. Bonß; H. Hartmann (Hg.): Entzauberte Wissenschaft. Zur Realität und Geltung soziologischer Forschung. Göttingen: Schwartz. S. 299-320.

Bergmann, Jörg; Meier, Christoph (2000). "Elektronische Prozessdaten und ihre Analyse." In: U. Flick; E. von Kardorff; I. Steinke (Hg.): Qualitative Forschung: Ein Handbuch. Reinbek: Rowohlt. S. 429-437.

Berners-Lee, Tim; Cailliau, Robert (1990): WorldWideWeb: Proposal for a HyperText Project. <http://www.w3.org/Proposal.html>, (28.08.2006)..

Berners-Lee, Tim; Fischetti, Mark (2000): Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor. Harper Business: New York, NY.

Blossfeld, Hans-Peter; Rohwer, Götz (2002): Techniques of event history modeling: new approaches to causal analysis. Erlbaum: Mahwah, NJ.

Blumenberg, Hans (1997): Schiffbruch mit Zuschauer: Paradigma einer Daseinsmetapher. Suhrkamp: Frankfurt am Main.

Bohnsack, Ralf (1999): Rekonstruktive Sozialforschung: Einführung in Methodologie und Praxis qualitativer Forschung. Leske + Budrich: Opladen.

Bönsch, Manfred (2000): Variable Lernwege: ein Lehrbuch der Unterrichtsmethoden. Schöningh: Zürich.

Boren, Ted M.; Ramey, Judith (2000): "Thinking Aloud: Reconciling Theory and Practice". IEEE Transactions on Professional Communication, 43, 3. S. 261-278.

Bortz, Jürgen (1999): Statistik für Sozialwissenschaftler. Springer: Heidelberg.

Böttger, A. (2001). "'Da haben wir richtig Mist gemacht. Zu Beginn und Ende 'devianter Sequenzen' in den Lebensgeschichten Jugendlicher." In: R. Sackmann; M. Wingens (Hg.): Strukturen des Lebenslaufs: Übergang - Sequenz – Verlauf. Weinheim: Juventa-Verlag. S. 51-77.

Brinkmann, Dieter (2000): Moderne Lernformen und Lerntechniken in der Erwachsenenbildung: Formen selbstgesteuerten Lernens. Bielefeld: Institut für Freizeitwissenschaft und Kulturarbeit (IFKA).

Brouns, Francis / Koper, Rob / Manderveld, Jocelyn / Bruggen, Jan van / Sloep, Peter / Rosmalen, Peter van / Tattersall, Colin / Vogten, Hubert (2005): A first exploration of an inductive analysis approach for detecting learning design patterns. Journal of Interactive Media in Education, 3.

Brüderl, J.; Klein, T. (2003). "Die Pluralisierung partnerschaftlicher Lebensformen im Kohortenvergleich." In: W. Bien; J. Marbach (Hg.): Partnerschaft und Familiengründung: Ergebnisse der dritten Welle des Familien-Survey. Opladen: Leske + Budrich.

Brüderl, J. (2004): "Die Pluralisierung partnerschaftlicher Lebensformen in Westdeutschland und Europa". Aus Politik und Zeitgeschichte, 19. S. 3-10.

Brüderl, Josef; Scherer, Stefani (2005). "Methoden zur Analyse von Sequenzdaten." In: A. Diekmann (Hg.): Methoden der Sozialforschung. VS Verlag für Sozialwissenschaften: Wiesbaden. S. 330-347.

Cadez, Igor; Heckerman, David; Smyth, Padhraic; White, Steven (2000): Visualization of Navigation Patterns on a Web Site using Model Based Clustering. Irvine.

Chakrabarti, Soumen (2000): "Data Mining for hypertext: A tutorial survey". SIGKDD Explorations, 1, 2.

Chakrabarti, Soumen (2003): Mining the Web: discovering knowledge from hypertext data. Morgan Kaufmann: Amsterdam.

Chan, Tak Wing (1995): "Optimal Matching Analysis: A methodological note on studying career mobility". Work and Occupation, 4. S. 467-490.

Chan, Tak Wing (1999): "Optimal Matching Analysis". Social Research Update, 24.

Chen, Ming-Syan; Han, Jiawei; Yu, Philip S. (1996): "Data Mining: An Overview from Database Perspective". Transactions on Knowledge and Data Engineering, 6, 8. S. 866-883.

Chen, Jiyang; Sun, Lisheng; Zaïane, Osmar R.; Goebel, Randy (2004). "Vizualizing and Discovering Web Navigational Patterns." In: Proceedings of the 7th International Workshop on the Web and Databases: Colocated with ACM SIGMOD/PODS 2004 (WebDB 04): Paris, France, June 17 - 18, 2004.

Clote, Peter; Backofen, Rolf (2002): Computational molecular biology: an introduction. Wiley: Weinheim.

Cooley, R.; Mobasher, B.; Srivastava, J. (1997). "Web Mining: Information and Pattern Discovery on the World Wide Web." In: Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).

Cooley, Robert; Tan, Pang-Ning, Srivastava, Jaideep (1999): Discovery of intersting usage patterns form web data. University of Minnesota, Technical Report TR99-022.

Delors, Jacques (1998): Learning, the treasure within. UNESCO Publishing: Paris.

Delors, Jacques (2004): "Toward lifelong education for all". In: Jérôme Bindé (Hg.): The future of values. 21st Century Talks. UNESCO Publishing: Paris. S. 181-186.

Denzin, Norman K. (1970): The research act: a theoretical introduction to sociological methods. Aldine: Chicago, Illinois.

Denzin, Norman K.; Lincoln, Yvonna S. (1994): Handbook of qualitative research. Sage: Thousand Oaks, California.

Dieberger, Andreas (2003): Social Connotations of Space in the Design for Virtual Communities and Social Navigation. In: Höök, K. / Benyon, D. (Hg.): Designing Information Spaces: The Social Navigation Approach. London: Springer.

Diekmann, Andreas (2007): Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen. Rowohlt: Reinbek.

Dijkstra, Wil; Toon, Taris (1995): "Measuring the Agreement between Sequences". Sociological Methods & Research, 4, 29. S. 532-535.

Dijkstra, Wil (2001): "How to Measure the Agreement Between Sequences: A Comment". Sociological Methods & Research, 29. S. 532-535.

Dohmen, Günther (1998): Zur Zukunft der Weiterbildung in Europa: lebenslanges Lernen für Alle in veränderten Lernumwelten. Bundesministerium für Bildung und Forschung: Bonn.

Dourish, P.; Chalmers, M. (1994): Running out of space: Models of information navigation. Short paper presented at HCI'94 (Glasgow, Scotland).

Driel, Kees van; Oosterveld, Paul (2001): "Nonoptimal Alignment: A Comment on 'Measuring the Agreement Between Sequences' by Dijkstra and Taris". Sociological Methods & Research, 29. S. 524-531.

Drummond, Kent; Hopper, Robert (1993): "Back Channels revisited: Acknowledgement tokens and speakership incipiency". Ressource on Language and Social Interaction, 26, 2.

Duncan, S. D.; Fiske, D. W. (1977): Face-to-Face interaction: Research, methods, and theory. Lawrence Earlbaum Associates: Hillsdale, N.Y.

Ehlers, Ulf-Daniel (2004): Qualität im E-Learning aus Lernersicht: Grundlagen, Empirie und Modellkonzeption subjektiver Qualität. Verlag für Sozialwissenschaften: Wiesbaden.

Elder, Glen H. (1985): Life course dynamics: trajectories and transitions, 1968 - 1980. Cornell University Press: Ithaca.

Elzinga, Cees (2003): "Sequence Similarity: A Nonaligning Technique". Sociological Methods & Research, 32. S. 3-29.

Elzinga, Cees (2005): "Combinatorial Representations of Token Sequences". Journal of Classification, 22, 1. S. 87-118.

Elzinga, Cees. (2005a): User Manual to Combinatorial Sequence Analyzer (CSA). Frije Universiteit Amsterdam.

Engelbart, Douglas C. (1963). "A Conceptual Framework for the Augmention of Man's Intellect." In: P. W. Howerton (Hg.): Vistas in Information Handling. Washington, D.C.: Spartan Books. S. 1-29.

Ericsson, K. Anders; Simon, Herbert A. (1980): "Verbal Report as Data". Psychological Review, 87, 3. S. 215-251.

Ericsson, Karl Anders; Simon, Herbert Alexander (1999): Protocol analysis: verbal reports as data. MIT Press: Cambridge (3, Revised Edition, Erstauflage 1984).

Erzberger, Christian; Prein, Gerald (1997): "Optimal-Matching-Technik: Ein Analyseverfahren zur Vergleichbarkeit und Ordnung individuell differenter Lebensverläufe". ZUMA-Nachrichten, 40, 21.

Erzberger, Christian (2001). "Sequenzmusteranalyse als fallorientierte Analysestrategie." In: R. Sackmann; M. Wingens (Hg.): Strukturen des Lebenslaufs: Übergang - Sequenz - Verlauf. Weinheim: Juventa. S. 135-162.

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases: an Overview." In: Advances in Knowledge Discovery an Data Mining: Menlo Park: AAAI Press. S. 1-34.

Feierabend, Sabine; Rathgeb, Thomas (2005): JIM-Studie 2005. Jugend, Information, (Multi) Media. Basisuntersuchung zum Medienumgang 12 - 19 jähriger. Medienpädagogischer Forschungsverbund Südwest: Stuttgart.

Flechsig, Karl-Heinz (1983): Der Göttinger Katalog Didaktischer Modelle: theoretische und methodologische Grundlagen. Zentrum für didaktische Studien: Nörten-Hardenberg.

Flechsig, Karl-Heinz (1996): Kleines Handbuch didaktischer Modelle. Neuland - Verlag für lebendiges Lernen: Eichenzell.

Flick, Uwe (2004): Triangulation: eine Einführung. Wiesbaden: VS Verlag für Sozialwissenschaften: Wiesbaden.

Fraley, C.; Raftery, A. (1998): "How many clusters? Which clustering method? Answers via model-based cluster analysis". Computer Journal, 41. S. 578-588.

Gusfield, Dan (1999): Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge University Press: Cambridge.

Häder, Michael (2006): Empirische Sozialforschung: eine Einführung. Verlag für Sozialwissenschaften: Wiesbaden.

Hagen, Christine; Niemann, Heike (2001). "Sozialhilfe als Sequenz im Lebenslauf. Institutionelle und individuelle Bedeutung der Übergänge aus der Sozialhilfe." In: R. Sackmann; M. Wingens (Hg.): Strukturen des Lebenslaufs: Übergang - Sequenz - Verlauf. Weinheim: Juventa-Verlag. S. 77-103.

Haller, Hans-Dieter (1995). "Wissensorganisation mit CEWID, einem wissensortierenten und tätigkeitsunterstützendem System." In: N. Meder (Hg.): Konstruktion und Retrieval von Wissen: 3. Tagung der Deutschen ISKO Sektion. Frankfurt / Main: Indeks-Verlag. S. 14-21.

Halpin, B.; Chan, T. W. (1998): "Class Careers as Sequences: An Optimal Matching Analysis of Work-Life Histories". European Sociological Review, 17, 2. S. 119-144.

Hamming, R. W. (1950): "Error-Detecting and Error-Correcting". Bell System Technical Journal, 2. S. 147-160.

Han, Shin-Kap; Moen, Phyllis (1999): "Clocking Out: Temporal Patterning of Retirement". The American Journal of Sociology, 1, 105. S. 191-236.

Hay, Birgit;Wets, Geert; Vanhoof, Koen (2001): "Clustering navigation patterns on a website using sequence alignment method". IJACAI's Workshop on Intelligent Techniques for Web Personalization.

Hay, Birgit; Wets, Geert; Vanhoof, Koen (2002). "Web Usage Mining by Means of Multidimensional Sequence Alignment Methods." In: O. R. Zaiane; J. Srivastava; M. Spiliopoulou; B. Masand (Hg.): Mining web data for discovering usage patterns and profiles: 4th international workshop, Edmonton, Canada, July 23, 2002. Springer: Berlin . S. 50-65.

Heimann, Paul (1976): Didaktik als Unterrichtswissenschaft. Klett: Stuttgart.

Hönigswald, Richard (1913): Studien zur Theorie pädagogischer Grundbegriffe: eine kritische Untersuchung. Spemann: Stuttgart.

Hönigswald, Richard (1927): Über die Grundlagen der Pädagogik: ein Beitrag zur Frage des pädagogischen Universitäts-Unterrichts. Reinhardt: München.

Höök, Kristina; Benyon, David; Munro, Alan J. (2003): Designing Information Spaces: The Social Navigation Approach. Springer: London.

Hooker, Giles; Finkelman, Matthew (2004): "Sequential Analysis for Learning Modes of Browsing." In: B. Mobasher; B. Liu; B. Masand; O. Nasraoui (Hg.): WebKDD 2004: Web Mining and Web Usage Analysis. Seattle, Washington.

Huber, Günter L.; Mandl, Heinz (1994): Verbale Daten: eine Einführung in die Grundlagen und Methoden der Erhebung und Auswertung. Beltz, Psychologie-Verlags-Union: Weinheim.

Iske, Stefan (2002): Vernetztes Wissen: Hypertext-Strategien im Internet. Bertelsmann: Bielefeld.

Iske, Stefan; Swertz, Christian. (2005): Methodologische Fragen der Verwendung von Bild-, Ton- und Textdaten zur Navigationsanalyse. <www.medienpaed.com/04-1/iske_swertz04-1.pdf>, (28.08.2006).

Jefferson, Gail (1993): "Caveat Speaker: Preliminary Notes on Recipient Topic-Shift Implicature". Research on Language and Social Interaction, 26, 1. S. 1-30.

Jeschke, Sabina; Keil-Slawik, Reinhard (2004). "Next Generation in eLearning Technology: Vom 'Typographischen Objekt' zum 'Ausführbaren Prozess'." GML 2004 - Grundfragen multimedialen Lehrens und Lernens. Alcatel SEL, Berlin, Germany.

Kaufman, Leonard; Rousseeuw, Peter J. (2005): Finding groups in data: an introduction to cluster analysis. Wiley: New York.

Koper, Rob (2006): Current Research in Learning Design. Educational Technology & Society, 9, 1. S. 13-22.

Kruskal, Joseph B. (1999). "An overview of sequence comparison." In: Sankoff, David; Kruskal, Joseph (Hg.): Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Stanford, California: CSLI. S. 1-44. (Erstauflage 1983).

Kuhlen, Rainer (1991): Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank. Springer: Berlin.

Landow, George P. (1997): Hypertext 2.0: Hypertext - the convergence of contemporary critical theory and technology. Johns Hopkins Univiversity Press: Baltimore, Md.

Lengrand, Paul (1972): Permanente Erziehung: eine Einführung. Verlag Dokumentation: München-Pullach.

Lengrand, Paul (1975): An introduction to lifelong education. Croom Helm: London.

Lengrand, Paul (1987): Areas of learning basic to lifelong education. UNESCO Institute for Education: Hamburg.

Leutner, Detlev (1992): Adaptive Lehrsysteme: instruktionspsychologische Grundlagen und experimentelle Analysen. PVU: Weinheim.

Levine, Joel H. (2000): "But what have you done for us lately? Commentary on Abbott and Tsay“. Sociological Methods & Research, 29, 1. S. 34-40.

Lewin, Kurt (1982): Feldtheorie. Huber: Bern.

Livingstone, Sonia; Bober, Magdalena (2005): UK Children Go Online. Final Report of key project findings. London School of Economics and Political Science, Department of Media and Communications.

Lu, Lin;Dunham, Margaret; Meng, Yu. (2005): Discovery of Significant Usage Patterns from Clusters of Clickstream Data. <http://db.cs.ualberta.ca/webkdd05/proc/paper4-Lu-Dunham-Meng.pdf>, (28.08.2006).

McLuhan, Marshall (1995): Die magischen Kanäle. Verlag der Kunst: Dresden.

Manovich, Lev (2001): The Language of New Media. MIT Press: Cambridge, Mass.

Markov, Andrei Andrejewitsch (1912): Wahrscheinlichkeitsrechnung. Teubner: Leipzig.

Meder, Norbert (1987): Der Sprachspieler: der postmoderne Mensch oder das Bildungsideal im Zeitalter der neuen Technologien. Janus-Presse: Köln.

Meder, Norbert (1995): Konstruktion und Retrieval von Wissen: 3. Tagung der Deutschen ISKO Sektion einschließlich der Vorträge des Workshops "Thesauri als Terminologische Lexika", Weilburg, 27. - 29.10.1993. Indeks-Verlag: Frankfurt / Main.

Meder, Norbert (1995a): Die Abbildung von Sachverhalten in die Zeit. Reflexion auf die philosophische Grundlegung der Pädagogik und die pädagogische Grundlegung der Erkenntnistheorie.

Meder, Norbert (1995b). "Didaktische Überlegungen zu einem veränderten Unterricht durch den Einsatz neuer Technologien." In: J. Lauffer; I. Vollkmer (Hg.): Kommunikative Kompetenz in einer sich verändernden Medienwelt. Leske + Budrich: Opladen. S. 48-63.

Meder, Norbert (1998): Neue Technologien und Erziehung / Bildung. Schneider-Verlag Hohengehren: Baltmannsweiler.

Meder, Norbert (2003). "Didaktische Anforderungen an Lernumgebungen." In: Ulf Ehlers (Hg.): E-Learning-Services im Spannungsfeld von Pädagogik, Ökonomie und Technologie. L3-lebenslanges Lernen im Bildungsnetzwerk der Zukunft. Bielefeld: Bertelsmann. S. 50-69.

Meder, Norbert (2004): Der Sprachspieler: der postmoderne Mensch oder das Bildungsideal im Zeitalter der neuen Technologien. Königshausen Neumann: Würzburg. (2., vollständig überarbeitete Auflage).

Meder, Norbert (2006): Web-Didaktik: eine neue Didaktik webbasierten vernetzten Lernens. Bertelsmann: Bielefeld.

Merrill, M. David; Tennyson, Robert D.; Posey, Larry O. (1992): Teaching concepts: an instructional design guide. Educational Technology Publications: Englewood Cliffs, NJ.

Merrill, M. David (1994): Instructional design theory. Educational Technology Publications: Englewood Cliffs, NJ.

Micheel, Heinz-Günther (2002): Explorative Dimensionierung und Typisierung von Rating-Skalen. Eine anwendungsorientierte Problembeschreibung. Universität Bielefeld, Fakultät für Pädagogik, (Habilitationsschrift).

Mobasher, B. ; Liu, B. ; Masand, B.; Nasraoui, O. (2004): WebKDD 2004: Web Mining and Web Usage Analysis. Proceedings of the Sixth International Workshop on Knowledge Discovery from the Web Seattle, Washington.

Mowitz-Lambert, Joachim (2001). "Übergangsmuster in der Statuspassage von beruflicher Ausbildung in den Erwerbsverlauf." In: R. Sackmann; M. Wingens (Hg.): Strukturen des Lebenslaufs: Übergang - Sequenz - Verlauf. Weinheim: Juventa-Verlag.

Nelson, Theodor (1965). "Complex information processing: a file structure for the complex, changing and the indeterminate." In: Association for Computing Machinery (ACM) (Hg.): Proceedings of the 1965 20th national conference. Cleveland, Ohio, United States: S. 84-100.

Olson, G. M.; Duffy, S. A.; Mack, R. L. (1984). "Thinking-out loud as a method for studying real-time comprehension processes." In: D. E. Kieras; M. A. Just (Hg.): New methods in reading comprehension research. Erlbaum: Hillsdale. S. 253-286.

Oostendorp, Herre van; Mul, Sjaak de (1999): "Learning by exploration: Thinking aloud while exploring an information system". Instructional Science, 27. S. 269-284.

Oyanagi, Shigeru; Kubota, Kazuto; Nakase, Akihiko (2002). "Mining WWW Access Sequence by Matrix Clustering." In: O. R. Zaiane; J. Srivastava; M. Spiliopoulou; B. Masand (Hg.): Mining web data for discovering usage patterns and profiles: 4th international workshop, Edmonton, Canada, July 23, 2002. Berlin: Springer.

Priemer, Burkhard (2004): Logfile-Analysen: Möglichkeiten und Grenzen ihrer Nutzung bei Untersuchungen der Mensch-Maschine-Interaktion. <www.medienpaed.com>, (28.08.2006).

Reigeluth, Charles M. (1983): Instructional-design theories and mode: An overview of their current status. Erlbaum: Hillsdale, NJ.

Reigeluth, Charles M. (1999): A new paradigm of instructional theory. Erlbaum: Mahwah, NJ.

Rohwer, Götz; Pötter, Ulrich (2002): Methoden sozialwissenschaftlicher Datenkonstruktion. Juventa: Weinheim.

Sacks, Harvey; Schegloff, Emanuel A.; Jefferson, Gail (1974): "A simplest systematics for the organisation of turn-taking in conversation". Language, 50, 4. S. 695-735.

Sackmann, Reinhold; Wingens, Matthias (2001). "Theoretische Konzepte des Lebenslaufs: Übergang, Sequenz und Verlauf." In: R. Sackmann; M. Wingens (Hg.): Strukturen des Lebenslaufs: Übergang - Sequenz - Verlauf. Weinheim: Juventa-Verlag. S. 17-48.

Sankoff, David; Kruskal, Joseph (1999): Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Stanford, California: CSLI. S. 1-44. (Erstauflage 1983).

Schegloff, Emanuel A. (1982). "Discourse as an interactional achivement: some uses of 'uh huh' and other things that come between sentences." In: D. Tannen (Hg.): Analyzing discourse: text and talk. Washington D.C.: Georgetown University Press. S. 71-93.

Scherer, Stefani (1999): "Early Career Patterns - a Comparison of Great Britain and West Germany". Mannheim Centre for European Social Research (MZES): Working Papers, 7.

Schnell, Rainer;Hill, Paul Bernhard; Esser, Elke (2005): Methoden der empirischen Sozialforschung. Oldenbourg: München.

Srivastava, Jaideep; Cooley, Robert; Deshpande, Mukund; Tan, Pan-Ning (1999): "Web Usage Mining: Discovery and Application of Usage Patterns from Web Data". SIGKDD Explorations, 1, 2.

Stovel, Katherine (2001): "Local Sequential Patterns: The Structure of Lynching in the Deep South, 1882 - 1930". Sociological Methods & Research, 79. S. 843-880.

Stovel, Katherine; Bolan, Marc (2004): "Residential Trajectories: Using optimal alignement to reveal the structure of residential mobility". Sociological Methods & Research, 32, 44. S. 559-598.

Swertz, Christian (2001): Computer und Bildung: eine medienanalytische Untersuchung der Computertechnologie in bildungstheoretischer Perspektive. Universität Bielefeld, Dissertation.

Swertz, Christian (2003): Didaktische Aufbereitung von Lernmaterialien. Universität Rostock., Zentrale Verwaltung Dezernat Studium und Lehre: Rostock.

Swertz, Christian (2004): Didaktisches Design: ein Leitfaden für den Aufbau hypermedialer Lernsysteme mit der Web-Didaktik. Bertelsmann: Bielefeld.

Swertz, Christian (2004a). "Selbstevaluation im Online-Lernen." In: D. M. Meister (Hg.): Online-Lernen und Weiterbildung. Verlag für Sozialwissenschaften: Wiesbaden. S. 177-189.

Taris, Toon (2000): A primer in longitudinal data analysis. Sage Publications: London.

Trautner, Hanns Martin (1992): Lehrbuch der Entwicklungspsychologie Bd. 1: Grundlagen. Hogrefe: Göttingen.

van Eimeren, Birgit; Frees, Beate (2005): "Nach dem Boom: Größter Internetzuwachs in bildungsfernen Gruppen". ARD-ZDF-Online-Studie 2005. Media Perspektiven, 8.

Wagner, R. A. (1999). "On the Complexity of the Extended String-to-String Correction Problem." In: D. Sankoff; J. Kruskal (Hg.): Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Stanford, Calif.: CSLI. S. 215-236. (Erstauflage 1983).

Wagner, Frank (2003): (N)ONLINER Atlas 2003: eine Topographie des digitalen Grabens durch Deutschland. TNSEmnid / Initiative D21: Berlin.

Wang, Weinan; Zaïane, Osmar R. (2002). "Clustering Web Sessions by Sequence Alignment." Third International Workshop on Management of Information on the Web, Aix en Provence, France. S. 394-398.

Weber, Max (1990): Wirtschaft und Gesellschaft: Grundriß der verstehenden Soziologie. Mohr: Tübingen. (Erstauflage 1922).

Webb, Eugene John;Campbell, Donald T.; Schwartz, Richard D.; Sechrest, Lee (1966): Unobtrusive measures: nonreactive research in the social sciences. Rand McNally: Chicago.

Windzio, Michael (2001). "Übergänge und Sequenzen. Der Einfluss von Arbeitslosigkeit auf den weiteren Erwerbsverlauf." In: R. Sackmann; M. Wingens (Hg.): Strukturen des Lebenslaufs: Übergang - Sequenz - Verlauf. Juventa-Verlag: Weinheim.

Wishart, David (1984): CLUSTAN-Benutzerhandbuch. Fischer: New York.

Wishart, David (1999): ClustanGraphics Primer: A Guide to Cluster Analysis. Edinburg: Clustan Ltd.

Wu, Lawrence L. (2000): "Some Comments on 'Sequence Analysis and Optikal Matching Methods in Sociology: Review and Prospekt'". Sociological Methods & Research, 29, 1. S. 41-64.

Yngve, Victor H. (1970). "On getting a word in edgewise." In: Department of Linguistics (Hg.): Papers from the sixth regional meeting of the Chicago Linguistic Society. University of Chicago. S. 567-578.

Ypma, Alexander; Heskes, Tom (2002). "Automatic Categorizaton of Web Pages and User Clustering with Muxures of Hidden Markov Models." In: O. R. Zaiane; J. Srivastava; M. Spiliopoulou; B. Masand (Hg.): Mining web data for discovering usage patterns and profiles: 4th international workshop, Edmonton, Canada, July 23, 2002. Springer: Berlin. S. 35-49.

Zaïane, Osmar R.; Xin, Man; Han, Jiawei (1998). "Discovering web access patterns and trends by applying OLAP and data mining technology on web logs." Proceedings Advances in Digital Libraries ADL'98. Santa Barbara. S. 19-29.

Zaïane, Osmar R.; Luo, Jun (2001). "Towards Evaluating Learners' Behavior in a Web-Based Distance Learning Environment." In: Proceedings of IEEE International Conference on Advanced Learning Technologies (ICAL 01), Madison. S. 357-360.

Zaïane, Osmar R. (2001). "Web Usage Mining for a better web-based learning environment." In: Proceedings of the Conference on Advanced Technology for Education. Banff, Alberta. S. 60-64.

Zaïane, Osmar R. (2002): "Building a Recommender Agent for e-Learning Systems". Proceedings of the International Conference on Computers in Education, S. 55-59.

15 Abbildungsverzeichnis

16 Tabellenverzeichnis

17 Anhang

17.1 Clusteralgorithmen

Beschreibung unterschiedlicher Verfahren der Clusteranalyse (Auszug aus Wishart 1999: 13f.).

17.1.1 „Single Linkage Method

The similarity between two clusters is defined as the proximity between the two most similar cases, one case from each cluster.

It is common, therefore, for clusters to be long and straggly when formed by single linkage, as a chain of links is easily established which links cases together. This is known as “chaining”, and can be particularly evident in large samples.

Single link clusters are isolated but need not be cohesive.

This method is sometimes referred to as the “Minimum Method” or “Nearest Neighbour”, because the union of two clusters is determined by those two members which are their nearest neighbours. Contrast with complete linkage.

17.1.2 Complete Linkage Method

The similarity between two clusters is defined as the proximity between the two most dissimilar cases, one case from each cluster.

Clusters formed by complete linkage tend to be spherically shaped because, in geometrical terms, the fusion distance corresponds to the cluster’s diameter. It has the characteristic that all cases within a cluster are mutually similar at the fusion level at which the cluster is created.

Complete link clusters are cohesive by definition, but need not be isolated.

This method is sometimes referred to as the “Maximum Method” or “Furthest Neighbour”, because the union of two clusters is determined by those two members which are furthest apart - their furthest neighbours. Contrast with single linkage.

17.1.3 Average linkage Method

The similarity between two clusters is defined as the average of all proximity values between pairs of cases, one case from each cluster.

The essence of this method is a compromise between the extremes of single linkage and complete linkage. Average linkage has the merit that the distribution of all members within two clusters influences the proximity between the clusters. Hence an outlier has less influence than for complete linkage; and the long straggly clusters formed by single linkage are discouraged.

Average Linkage seeks partitions which minimise the sum of the average within-cluster dissimilarities, or maximise the sum of the average within-cluster similarities.

This method has also been referred to as the Unweighted Pair Group Method using Arithmetic Averages (UPGMA).

17.1.4 Weighted Average Linkage Method

When two clusters p and q are combined, the similarity Sr,pq between any other cluster r and the newly formed cluster p+q is the simple average of Srp and Srq thus:

Sr,pq = ½{ Srp + Srq}

This method has been described as weighting the member most recently admitted to a cluster equal with all previous members. It has also been referred to as the Weighted Pair Group Method using Arithmetic Averages (WPGMA).

17.1.5 Centroid Method

The Centroid method is only defined in terms of squared distances. The squared distance between two clusters is defined as the squared distance between the cluster means, or centroids. The size or weight of a cluster is not relevant, though its spatial distribution is used in the calculation of the centroid.

This method should, strictly speaking, only be used with a matrix of squared distances. However, its usage has been extended to dissimilarities; and although ClustanGraphics allows Centroid to be used with similarities, the resulting fusion values are not theoretically defined. See also converting similarities to dissimilarities.

Centroid can exhibit tree reversals. These arise because the new squared distance Dr,pq between any cluster r and the cluster formed by the union of clusters p and q can be less than Drp or Drq. It is simple to construct a geometrical example which illustrates this feature.

Centroid is sometimes referred to as the Unweighted Pair-Group Method of Clustering (UPGMC).

17.1.6 Mean Proximity Method

Mean Proximity maximises the average of the within-cluster similarities or minimises the average of the between-cluster dissimilarities, for all cluster comparisons.

Contrast with Average Linkage and Weighted Average Linkage clustering.

17.1.7 Median Method

The Median method is only defined in terms of squared distances. The new squared distance Dr,pq between any cluster r and the cluster formed by the union of clusters p and q is defined as the squared distance between the centroid of cluster r and the midpoint of the line between the centroids of clusters p and q. The size or weight of a cluster is not relevant to Dr,pq.

This method should, strictly speaking, only be used with a matrix of squared distances. However, its usage has been extended to dissimilarities; and although ClustanGraphics allows Median to be used with similarities, the resulting fusion values are not theoretically defined. See also converting similarities to dissimilarities.

Median can exhibit tree reversals. These arise because the new squared distance Dr,pq between any cluster r and the cluster formed by the union of clusters p and q can be less than Drp or Drq. It is simple to construct a geometrical example which illustrates this aspect.

Median is sometimes referred to as the Weighted Pair-Group Method of Clustering (WPGMC).

17.1.8 Increase in Sum of Squares (Ward's Method)

The dissimilarity between two clusters is defined as the increase in the sum of squares which would result from the union of the two clusters

The sum of squares function is only defined for squared distances. For a given partition of the sample, it is the sum of the squared distances between the cases and the centres (or means) of the clusters to which they belong. Increase in Sum of Squares seeks to minimise this function. In this respect it is very similar to the Sum of Squares method. Increase in Sum of Squares should, strictly speaking, only be used with a squared distance matrix. However, its usage has been extended to dissimilarities; and although ClustanGraphics allows it to be used with similarities, the resulting fusion values are not theoretically defined. See also converting similarities to dissimilarities.

This method is sometimes referred to as the Incremental Sum of Squares."

17.2 Levenshtein-Distanzen (*default*)

Die folgende Tabelle enthält die Levenshtein-Distanzen für den paarweisen Vergleich der Beispielsequenzen aus Kapitel 7: 61, *Sequenzanalyse am Beispiel* auf Grundlage der Default-Definition der Substitutionskosten.

Die erste Spalte enthält die Fallnummer der Ausgangssequenz; die zweite Spalte enthält die Fallnummer der Zielsequenz; die dritte Spalte enthält die Anzahl der Elemente der Ausgangssequenz; die vierte Spalte enthält die Anzahl der Elemente der Zielsequenz und die fünfte Spalte enthält die Levenshtein-Distanz.

2	1	5	5	0.00
3	1	5	5	0.00
3	2	5	5	0.00
4	1	5	5	2.00
4	2	5	5	2.00
4	3	5	5	2.00
5	1	5	5	2.00
5	2	5	5	2.00
5	3	5	5	2.00
5	4	5	5	0.00
6	1	5	5	2.00
6	2	5	5	2.00
6	3	5	5	2.00
6	4	5	5	0.00
6	5	5	5	0.00
7	1	5	5	4.00
7	2	5	5	4.00
7	3	5	5	4.00
7	4	5	5	4.00
7	5	5	5	4.00
7	6	5	5	4.00
8	1	5	5	4.00
8	2	5	5	4.00
8	3	5	5	4.00
8	4	5	5	4.00
8	5	5	5	4.00
8	6	5	5	4.00
8	7	5	5	0.00
9	1	5	5	4.00
9	2	5	5	4.00
9	3	5	5	4.00
9	4	5	5	4.00
9	5	5	5	4.00
9	6	5	5	4.00
9	7	5	5	0.00
9	8	5	5	0.00
10	1	5	5	8.00
10	2	5	5	8.00
10	3	5	5	8.00
10	4	5	5	6.00
10	5	5	5	6.00
10	6	5	5	6.00
10	7	5	5	4.00
10	8	5	5	4.00
10	9	5	5	4.00
11	1	5	5	4.00
11	2	5	5	4.00
11	3	5	5	4.00
11	4	5	5	4.00
11	5	5	5	4.00
11	6	5	5	4.00
11	7	5	5	4.00
11	8	5	5	4.00
11	9	5	5	4.00
11	10	5	5	4.00

17.3 Optimal-Matching „test output file“ (TDA)

Die folgende Darstellung enthält das gekürzte „test output file“ der Optimal-Matching Analyse mit TDA, das die Dokumentation der Indel- und Substitutionskosten enthält, sowie die Matrizen zur Errechnung der Levenshtein-Distanz.

Insgesamt umfasst das Ausgabedokument n (m-1) / 2 Sequenzvergleiche: für die Beispieldatei mit 11 Sequenzen ergeben sich demnach 55 Sequenzvergleiche. Da die Sequenzen 1, 2, 3 und 4,5,6 sowie 7,8,9 identisch sind, werden die daraus resultierenden identischen Sequenzvergleiche nicht dargestellt.

```
Optimal matching test output file.
Number of states: 5
Max sequence lenght: 5

Indel cost
1 1 1 1 1

Substitution cost
0 2 2 2 2
2 0 2 2 2
2 2 0 2 2
2 2 2 0 2
2 2 2 2 0

D Matrix        0       1       2       3       4       5
          ---------------------------------------------------
                B       1       2       3       4       5
          ---------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    1 |   1.00    0.00    1.00    2.00    3.00    4.00
   2    2 |   2.00    1.00    0.00    1.00    2.00    3.00
   3    3 |   3.00    2.00    1.00    0.00    1.00    2.00
   4    4 |   4.00    3.00    2.00    1.00    0.00    1.00
   5    5 |   5.00    4.00    3.00    2.00    1.00    0.00

[...}

D Matrix        0       1       2       3       4       5
          ---------------------------------------------------
                B       1       2       3       4       5
          ---------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    2 |   1.00    2.00    1.00    2.00    3.00    4.00
   2    1 |   2.00    1.00    2.00    3.00    4.00    5.00
   3    3 |   3.00    2.00    3.00    2.00    3.00    4.00
   4    4 |   4.00    3.00    4.00    3.00    2.00    3.00
   5    5 |   5.00    4.00    5.00    4.00    3.00    2.00

[...]

D Matrix        0       1       2       3       4       5
          ---------------------------------------------------
                B       2       1       3       4       5
          ---------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    2 |   1.00    0.00    1.00    2.00    3.00    4.00
   2    1 |   2.00    1.00    0.00    1.00    2.00    3.00
   3    3 |   3.00    2.00    1.00    0.00    1.00    2.00
   4    4 |   4.00    3.00    2.00    1.00    0.00    1.00
   5    5 |   5.00    4.00    3.00    2.00    1.00    0.00

[...]

D Matrix        0       1       2       3       4       5
          ---------------------------------------------------
                B       2       1       3       4       5
          ---------------------------------------------------
   0    A |   0.00    1.00    2.00    3.00    4.00    5.00
   1    2 |   1.00    0.00    1.00    2.00    3.00    4.00
```

```
   2   1 |   2.00    1.00    0.00    1.00    2.00    3.00
   3   3 |   3.00    2.00    1.00    0.00    1.00    2.00
   4   4 |   4.00    3.00    2.00    1.00    0.00    1.00
   5   5 |   5.00    4.00    3.00    2.00    1.00    0.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       2       3       4       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    0.00    1.00    2.00    3.00    4.00
   2   4 |   2.00    1.00    2.00    3.00    2.00    3.00
   3   3 |   3.00    2.00    3.00    2.00    3.00    4.00
   4   2 |   4.00    3.00    2.00    3.00    4.00    5.00
   5   5 |   5.00    4.00    3.00    4.00    5.00    4.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       4       3       2       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    0.00    1.00    2.00    3.00    4.00
   2   4 |   2.00    1.00    0.00    1.00    2.00    3.00
   3   3 |   3.00    2.00    1.00    0.00    1.00    2.00
   4   2 |   4.00    3.00    2.00    1.00    0.00    1.00
   5   5 |   5.00    4.00    3.00    2.00    1.00    0.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       2       1       3       4       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    2.00    1.00    2.00    3.00    4.00
   2   4 |   2.00    3.00    2.00    3.00    2.00    3.00
   3   3 |   3.00    4.00    3.00    2.00    3.00    4.00
   4   2 |   4.00    3.00    4.00    3.00    4.00    5.00
   5   5 |   5.00    4.00    5.00    4.00    5.00    4.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       4       3       2       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    0.00    1.00    2.00    3.00    4.00
   2   4 |   2.00    1.00    0.00    1.00    2.00    3.00
   3   3 |   3.00    2.00    1.00    0.00    1.00    2.00
   4   2 |   4.00    3.00    2.00    1.00    0.00    1.00
   5   5 |   5.00    4.00    3.00    2.00    1.00    0.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       2       3       4       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   5 |   1.00    2.00    3.00    4.00    5.00    4.00
   2   4 |   2.00    3.00    4.00    5.00    4.00    5.00
   3   3 |   3.00    4.00    5.00    4.00    5.00    6.00
   4   2 |   4.00    5.00    4.00    5.00    6.00    7.00
   5   1 |   5.00    4.00    5.00    6.00    7.00    8.00

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       2       3       4       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   5 |   1.00    2.00    3.00    4.00    5.00    4.00
   2   4 |   2.00    3.00    4.00    5.00    4.00    5.00
```

```
   3   3 |   3.00    4.00    5.00    4.00    5.00    6.00
   4   2 |   4.00    5.00    4.00    5.00    6.00    7.00
   5   1 |   5.00    4.00    5.00    6.00    7.00    8.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       2       1       3       4       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   5 |   1.00    2.00    3.00    4.00    5.00    4.00
   2   4 |   2.00    3.00    4.00    5.00    4.00    5.00
   3   3 |   3.00    4.00    5.00    4.00    5.00    6.00
   4   2 |   4.00    3.00    4.00    5.00    6.00    7.00
   5   1 |   5.00    4.00    3.00    4.00    5.00    6.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       4       3       2       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   5 |   1.00    2.00    3.00    4.00    5.00    4.00
   2   4 |   2.00    3.00    2.00    3.00    4.00    5.00
   3   3 |   3.00    4.00    3.00    2.00    3.00    4.00
   4   2 |   4.00    5.00    4.00    3.00    2.00    3.00
   5   1 |   5.00    4.00    5.00    4.00    3.00    4.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       2       3       4       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    0.00    1.00    2.00    3.00    4.00
   2   3 |   2.00    1.00    2.00    1.00    2.00    3.00
   3   5 |   3.00    2.00    3.00    2.00    3.00    2.00
   4   4 |   4.00    3.00    4.00    3.00    2.00    3.00
   5   2 |   5.00    4.00    3.00    4.00    3.00    4.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       2       1       3       4       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    2.00    1.00    2.00    3.00    4.00
   2   3 |   2.00    3.00    2.00    1.00    2.00    3.00
   3   5 |   3.00    4.00    3.00    2.00    3.00    2.00
   4   4 |   4.00    5.00    4.00    3.00    2.00    3.00
   5   2 |   5.00    4.00    5.00    4.00    3.00    4.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       1       4       3       2       5
          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    0.00    1.00    2.00    3.00    4.00
   2   3 |   2.00    1.00    2.00    1.00    2.00    3.00
   3   5 |   3.00    2.00    3.00    2.00    3.00    2.00
   4   4 |   4.00    3.00    2.00    3.00    4.00    3.00
   5   2 |   5.00    4.00    3.00    4.00    3.00    4.00

[...]

D Matrix        0       1       2       3       4       5
          ------------------------------------------------
                B       5       4       3       2       1

          ------------------------------------------------
   0   A |   0.00    1.00    2.00    3.00    4.00    5.00
   1   1 |   1.00    2.00    3.00    4.00    5.00    4.00
```

```
  2   3 |   2.00    3.00    4.00    3.00    4.00    5.00
  3   5 |   3.00    2.00    3.00    4.00    5.00    6.00
  4   4 |   4.00    3.00    2.00    3.00    4.00    5.00
  5   2 |   5.00    4.00    3.00    4.00    3.00    4.00
```

17.4 TDA Ausgabedatei (*.tst): *default*-Substitutionskosten

Darstellung der TDA-Ausgabedatei des beispielhaften Datensatzes (vgl. Kap. 7.1: 62) mit der Dokumentation der Durchführung der Optimal-Matching Analyse auf Grundlage der *default*-Definition der Substitutionskosten (vgl. Zeile 185, 186).

```
101.TDA. Analysis of Transition Data (6.4k). Sun May 21 22:52:21 2006
102.Current memory: 330832 bytes.
103.
104.Reading command file: D:\_Diss\Diss_Navigation\logdaten\grep\testwien\testwien.cf
105.================================================================================
106.rspss1(...)=D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq01.sav
107.Reading SPSS sav file: D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq01.sav
108.
109.Identification: $FL2@(#) SPSS DATA FILE MS Windows Release 12.0 spssio32.dll
110.Number of OBS elements per observation: 6
111.Compression switch: 1
112.Index of case-weight variable: 0
113.Number of cases: 11
114.Compression bias: 100
115.
116.Creation date: 19 May 06
117.Creation time: 16:27:34
118.File label:
119.
120.Number of variables: 6
121.Number of string variables: 0
122.
123.Reading data to check variables.
124.Read 11 records.
125.Number of blank-type missing values: 0
126.Number of system-type missing values: 0
127.
128.Idx Variable  T   S  PFmt  Definition
129.-------------------------------------
130.  1 NUM       3   1   2.0  spss(0)
131.  2 V1        3   1   2.0  spss(0)
132.  3 V2        3   1   2.0  spss(0)
133.  4 V3        3   1   2.0  spss(0)
134.  5 V4        3   1   2.0  spss(0)
135.  6 V5        3   1   2.0  spss(0)
136.
137.Reading data again to create internal data matrix.
138.Maximum number of cases: 11
139.Allocated 66 bytes for data matrix.
140.
141.Read 11 records.
142.Created a data matrix with 6 variables and 11 cases.
143.--------------------------------------------------------------------------------
144.nvar(...)
145.Creating new variables. Current memory: 331037 bytes.
146.
147.Idx Variable  T   S  PFmt  Definition
148.-------------------------------------
```

```
149.  1 ID        3   4   0.0  NUM
150.  2 Y0        3   4   0.0  V1
151.  3 Y1        3   4   0.0  V2
152.  4 Y2        3   4   0.0  V3
153.  5 Y3        3   4   0.0  V4
154.  6 Y4        3   4   0.0  V5
155.
156.New variables will be added to existing data matrix.
157.Trivial matching.
158.
159.Added 6 variable(s) to existing data matrix.
160.Number of cases with no match: 0
161.
162.End of creating new variables. Current memory: 331410 bytes.
163.--------------------------------------------------------------------------------
164.seqdef=Y0,,Y4
165.Creating a new sequence data structure. Current memory: 331410 bytes.
166.Sequence structure number: 1
167.Sequence type: 1
168.Currently defined sequences:
169.
170.Sequence          State       Time axis       Number
171.Structure Type  Variables  Minimum  Maximum  of States  States
172.-------------------------------------------------------------
173.     1      1         5          0        4         5    1 2 3 4 5
174.
175.Range of common time axis: 0 to 4.
176.--------------------------------------------------------------------------------
177.seqm(...)=D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq_default.df
178.Sequence proximity measures. Current memory: 331442 bytes.
179.Optimal matching.
180.Using sequence data structure 1.
181.Number of states: 5. Max sequence length: 5
182.Option (sm=2): skip identical states.
183.Test output will be written to: wienseq01_default_tst1.tst
184.
185.Default indel cost: 1.
186.Default substitution cost: 2.
187.
188.Starting alignment procedure.
189.Number of sequences (cases): 11
190.Sequences with zero length or internal gaps: 0
191.Sequences used for alignment: 11
192.
193.Number of alignments: 55
194.55             record(s)             written             to             output             file:
  D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq_default.df
195.Maximum distance between sequences 10 and 1: 8
196.TDA description written to: wienseq01_default_tst1.tda
197.--------------------------------------------------------------------------------
198.Current memory: 330832 bytes. Max memory used: 331766 bytes.
199.End of program. Sun May 21 22:52:21 2006
```

17.5 Dokumentation der Zugriffe auf die Lernumgebung

Die folgende Abbildung 75 dokumentiert die Zugriffe auf die Lernumgebung als Ausgangspunkt der Sequenzerstellung am Beispiel einer SPSS-Datei.

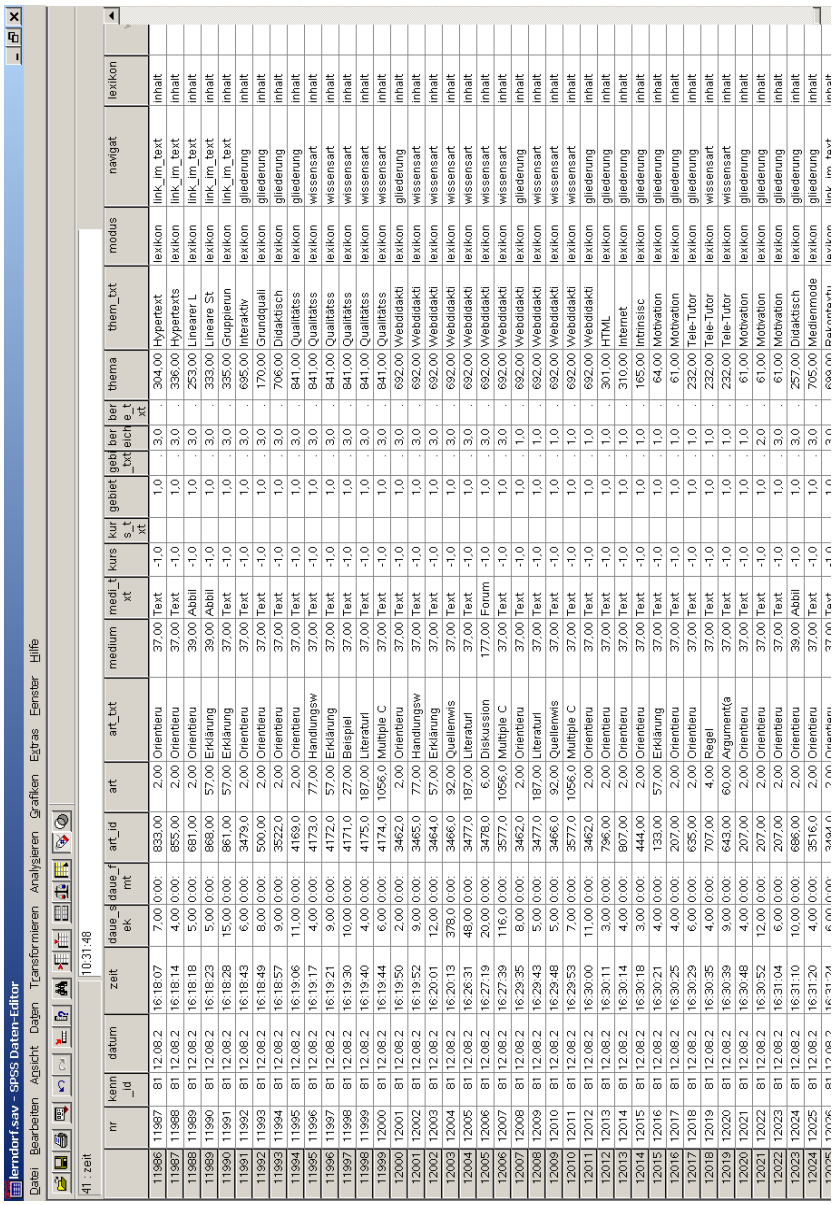

Abbildung 75: Dokumentation der Logdaten.

17.6 Dokumentation der erzeugten Sequenzen (Beispiel)

Abbildung 76 dokumentiert beispielhaft die Sequenzen der Zugriffen auf die Lernumgebung als Ausgangspunkt der Sequenzanalyse.

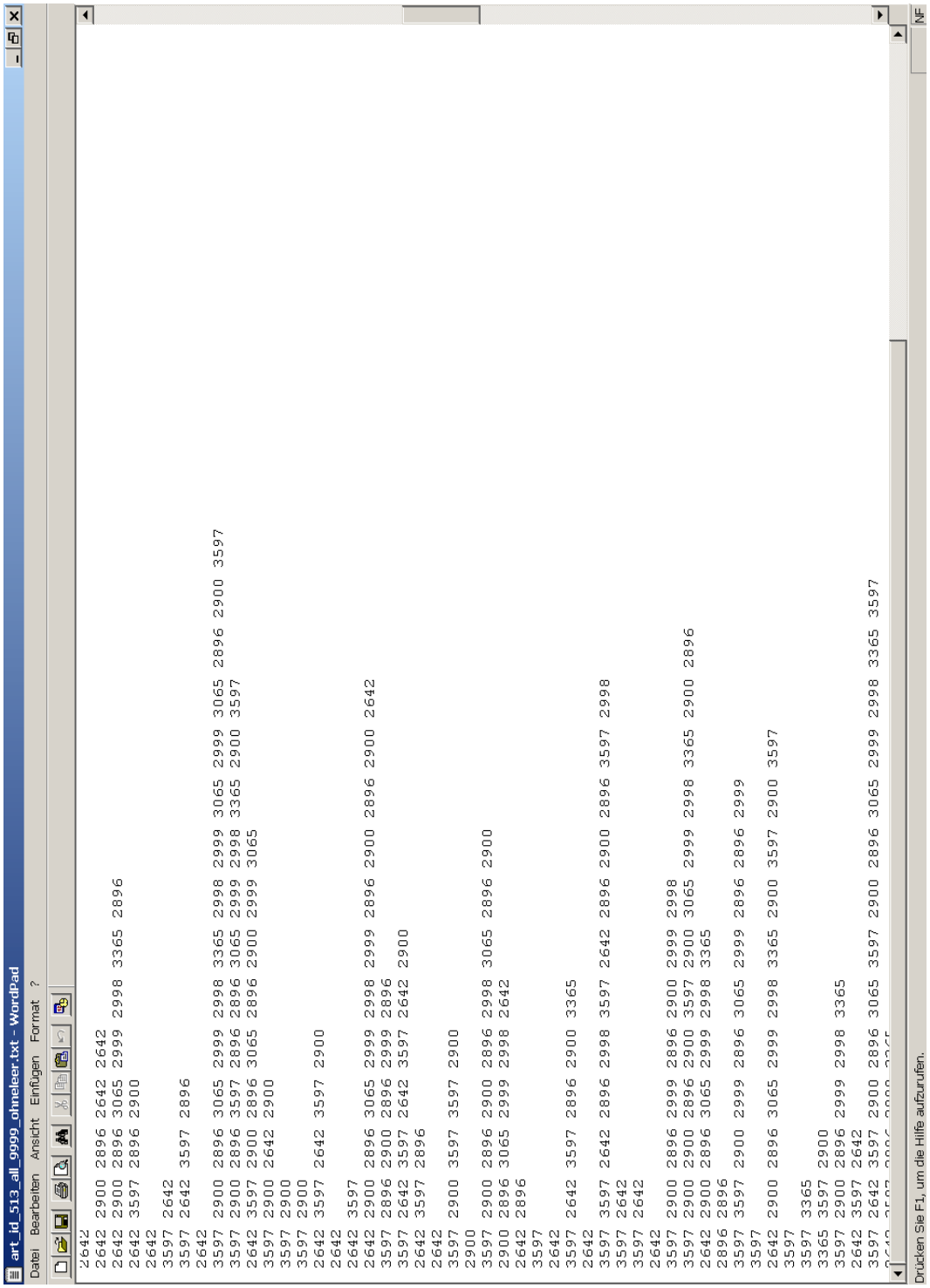

Abbildung 76: Sequenzen als Ausgangsdaten der Optimal-Matching Analyse.

17.7 Aggregierte Logfile-Analyse

Die folgende Grafik veranschaulicht die Funktionalität der Analyse aggregierter Logdaten am Beispiel der Logfile-Analysesoftware „Sawmill“ (<http://www.sawmill.net/>, (28.08.2006).

Abbildung 77: Aggregierte Logfile-Analyse mit 'Sawmill'.

17.8 TDA Ausgabedatei (*.tst): datenbasierte Substitutionskosten

Darstellung der TDA-Ausgabedatei des beispielhaften Datensatzes (vgl. Kap. 7.2: 74) mit der Dokumentation der Durchführung der Optimal-Matching Analyse auf Grundlage der datenbasierten Definition der Substitutionskosten (vgl. Zeile 86).

```
1. TDA. Analysis of Transition Data (6.4k). Tue Jul 18 23:14:00 2006
2. Current memory: 330832 bytes.
3.
4. Reading command file: D:\_Diss\Diss_Navigation\logdaten\grep\testwien\testwien.cf
5. ================================================================================
6. rspss1(...)=D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq01.sav
7. Reading SPSS sav file: D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq01.sav
8.
9. Identification: $FL2@(#) SPSS DATA FILE MS Windows Release 12.0 spssio32.dll
10.Number of OBS elements per observation: 6
11.Compression switch: 1
12.Index of case-weight variable: 0
13.Number of cases: 11
14.Compression bias: 100
15.
16.Creation date: 19 May 06
17.Creation time: 16:27:34
18.File label:
19.
20.Number of variables: 6
21.Number of string variables: 0
22.
23.Reading data to check variables.
24.Read 11 records.
25.Number of blank-type missing values: 0
26.Number of system-type missing values: 0
27.
28.Idx Variable  T   S  PFmt  Definition
29.-------------------------------------
30.  1 NUM       3   1   2.0  spss(0)
31.  2 V1        3   1   2.0  spss(0)
32.  3 V2        3   1   2.0  spss(0)
33.  4 V3        3   1   2.0  spss(0)
34.  5 V4        3   1   2.0  spss(0)
35.  6 V5        3   1   2.0  spss(0)
36.
37.Reading data again to create internal data matrix.
38.Maximum number of cases: 11
39.Allocated 66 bytes for data matrix.
40.
41.Read 11 records.
42.Created a data matrix with 6 variables and 11 cases.
43.------------------------------------------------------------------------------
44.nvar(...)
45.Creating new variables. Current memory: 331037 bytes.
46.
47.Idx Variable  T   S  PFmt  Definition
48.-------------------------------------
```

```
 1  ID        3   4   0.0  NUM
 2  Y0        3   4   0.0  V1
 3  Y1        3   4   0.0  V2
 4  Y2        3   4   0.0  V3
 5  Y3        3   4   0.0  V4
 6  Y4        3   4   0.0  V5

New variables will be added to existing data matrix.
Trivial matching.

Added 6 variable(s) to existing data matrix.
Number of cases with no match: 0

End of creating new variables. Current memory: 331410 bytes.
------------------------------------------------------------------------------
seqdef=Y0,,Y4
Creating a new sequence data structure. Current memory: 331410 bytes.
Sequence structure number: 1
Sequence type: 1
Currently defined sequences:

Sequence          State       Time axis       Number
Structure Type  Variables  Minimum  Maximum  of States  States
---------------------------------------------------------------
     1      1         5          0        4         5    1 2 3 4 5

Range of common time axis: 0 to 4.
------------------------------------------------------------------------------
seqm(...)=D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq_datenb.df
Sequence proximity measures. Current memory: 331442 bytes.
Optimal matching.
Using sequence data structure 1.
Number of states: 5. Max sequence length: 5
Option (sm=2): skip identical states.
Test output will be written to: wienseq01_datenb_tst3.tst

Indel cost: 1.
Substitution cost based on data, type 2.

Starting alignment procedure.
Number of sequences (cases): 11
Sequences with zero length or internal gaps: 0
Sequences used for alignment: 11

Number of alignments: 55
55             record(s)             written             to             output             file:
  D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq_datenb.df
Maximum distance between sequences 10 and 1: 6.90909
TDA description written to: wienseq01_datenb_tst3.tda
------------------------------------------------------------------------------
Current memory: 330832 bytes. Max memory used: 331766 bytes.
End of program. Tue Jul 18 23:14:00 2006
```

17.9 TDA Ausgabedatei (*.tst): Substitutionskosten als absolute Differenz

Darstellung der TDA-Ausgabedatei des beispielhaften Datensatzes (vgl. Kap. 7.3: 82) mit der Dokumentation der Durchführung der Optimal-Matching Analyse auf Grundlage der Substitutionskosten als absoluter Differenz (vgl. Zeile 184, 185).

```
100.TDA. Analysis of Transition Data (6.4k). Fri May 19 21:31:58 2006
101.Current memory: 330832 bytes.
102.
103.Reading command file: D:\_Diss\Diss_Navigation\logdaten\grep\testwien\testwien.cf
104.================================================================================
105.rspss1(...)=D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq01.sav
106.Reading SPSS sav file: D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq01.sav
107.
108.Identification: $FL2@(#) SPSS DATA FILE MS Windows Release 12.0 spssio32.dll
109.Number of OBS elements per observation: 6
110.Compression switch: 1
111.Index of case-weight variable: 0
112.Number of cases: 11
113.Compression bias: 100
114.
115.Creation date: 19 May 06
116.Creation time: 16:27:34
117.File label:
118.
119.Number of variables: 6
120.Number of string variables: 0
121.
122.Reading data to check variables.
123.Read 11 records.
124.Number of blank-type missing values: 0
125.Number of system-type missing values: 0
126.
127.Idx Variable  T   S  PFmt  Definition
128.-----------------------------------
129.  1 NUM       3   1   2.0  spss(0)
130.  2 V1        3   1   2.0  spss(0)
131.  3 V2        3   1   2.0  spss(0)
132.  4 V3        3   1   2.0  spss(0)
133.  5 V4        3   1   2.0  spss(0)
134.  6 V5        3   1   2.0  spss(0)
135.
136.Reading data again to create internal data matrix.
137.Maximum number of cases: 11
138.Allocated 66 bytes for data matrix.
139.
140.Read 11 records.
141.Created a data matrix with 6 variables and 11 cases.
142.--------------------------------------------------------------------------------
143.nvar(...)
144.Creating new variables. Current memory: 331037 bytes.
145.
```

```
Idx Variable  T   S  PFmt  Definition
-------------------------------------
  1 ID        3   4   0.0  NUM
  2 Y0        3   4   0.0  V1
  3 Y1        3   4   0.0  V2
  4 Y2        3   4   0.0  V3
  5 Y3        3   4   0.0  V4
  6 Y4        3   4   0.0  V5

New variables will be added to existing data matrix.
Trivial matching.

Added 6 variable(s) to existing data matrix.
Number of cases with no match: 0

End of creating new variables. Current memory: 331410 bytes.
------------------------------------------------------------------------------
seqdef=Y0,,Y4
Creating a new sequence data structure. Current memory: 331410 bytes.
Sequence structure number: 1
Sequence type: 1
Currently defined sequences:

Sequence          State       Time axis       Number
Structure Type  Variables  Minimum  Maximum  of States  States
---------------------------------------------------------------
     1      1         5          0        4         5    1 2 3 4 5

Range of common time axis: 0 to 4.
------------------------------------------------------------------------------
seqm(...)=D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq_scost1.df
Sequence proximity measures. Current memory: 331442 bytes.
Optimal matching.
Using sequence data structure 1.
Number of states: 5. Max sequence length: 5
Option (sm=2): skip identical states.
Test output will be written to: wienseq01_scost1.tst

Indel cost: 1.
Substitution cost defined by absolute difference.

Starting alignment procedure.
Number of sequences (cases): 11
Sequences with zero length or internal gaps: 0
Sequences used for alignment: 11

Number of alignments: 55
55 record(s) written to output file:
  D:\_Diss\Diss_Navigation\logdaten\grep\testwien\wienseq_scost1.df
Maximum distance between sequences 10 and 1: 8
------------------------------------------------------------------------------
Current memory: 330832 bytes. Max memory used: 331766 bytes.
End of program. Fri May 19 21:31:58 2006
```

17.10 Levenshtein-Distanz (absolute Differenz)

Tabelle 11 enthält die Levenshtein-Distanz für die Substitutionskosten als absoluter Differenz (scost=1), vgl. Kapitel 7: 61, *Sequenzanalyse am Beispiel.*

2	1	5	5	0.00
3	1	5	5	0.00
3	2	5	5	0.00
4	1	5	5	2.00
4	2	5	5	2.00
4	3	5	5	2.00
5	1	5	5	2.00
5	2	5	5	2.00
5	3	5	5	2.00
5	4	5	5	0.00
6	1	5	5	2.00
6	2	5	5	2.00
6	3	5	5	2.00
6	4	5	5	0.00
6	5	5	5	0.00
7	1	5	5	4.00
7	2	5	5	4.00
7	3	5	5	4.00
7	4	5	5	4.00
7	5	5	5	4.00
7	6	5	5	4.00
8	1	5	5	4.00
8	2	5	5	4.00
8	3	5	5	4.00
8	4	5	5	4.00
8	5	5	5	4.00
8	6	5	5	4.00
8	7	5	5	0.00
9	1	5	5	4.00
9	2	5	5	4.00
9	3	5	5	4.00
9	4	5	5	4.00
9	5	5	5	4.00
9	6	5	5	4.00
9	7	5	5	0.00
9	8	5	5	0.00
10	1	5	5	8.00
10	2	5	5	8.00

10	3	5	5	8.00
10	4	5	5	6.00
10	5	5	5	6.00
10	6	5	5	6.00
10	7	5	5	4.00
10	8	5	5	4.00
10	9	5	5	4.00
11	1	5	5	4.00
11	2	5	5	4.00
11	3	5	5	4.00
11	4	5	5	4.00
11	5	5	5	4.00
11	6	5	5	4.00
11	7	5	5	4.00
11	8	5	5	4.00
11	9	5	5	4.00
11	10	5	5	4.00

Tabelle 11: Levenshtein-Distanz auf Grundlage der Substitutionskosten als absoluter Differenz (scost=1)

17.11 Ontologie der rezeptiven Wissensarten in der Web-Didaktik

(vgl. Meder 2006)

1 Orientierungswissen (know what, know if)
- 1.1 Historie
- 1.2 Szenario
  - 1.2.1 Hypothetische Situtation
  - 1.2.2 Geschichte (Narration)
  - 1.2.3 Virtuelle Welt
- 1.3 Fakten
- 1.4 Zusammenfassung
- 1.5 Überblick

2 Handlungswissen (know-how)
- 2.1 Regel
- 2.2 Prozedur
  - 2.2.1 Aministrative Anleitung
  - 2.2.2 Bedienungsanleitung
  - 2.2.3 Soziale Norm
- 2.3 Checkliste
- 2.4 Prinzip
- 2.5 Strategie
- 2.6 Gesetz
- 2.7 Verordnung
- 2.8 Gesetzeskommentar

3 Erklärungswissen (know why)
- 3.1 Warum Erklärung (know why im engeren Sinne)
  - 3.1.1 Schluss
  - 3.1.2 Beweis
- 3.2 Was-Erklärung (know why im erklärenden Sinne)
  - 3.2.1 (Lehr) Satz / Theorem
  - 3.2.2 Beschreibung
  - 3.2.3 Definition
    - 3.2.3.1 Mathematische Definition
    - 3.2.3.2 Begriffliche Definition
- 3.3 Fallerklärung
  - 3.3.1 Beispiel
  - 3.3.2 Gegenbeispiel
- 3.4 Argument(ation)
- 3.5 Vermutung / Annahme
- 3.6 Hypothese

3.7 Reflexion
3.8 Erläuterung
3.9 Deutung / Interpretation

4 Quellenwissen (know where)
4.1 Archiv Referenz
4.1.1 Dokument Referenz
4.1.1.1 Statistik
4.1.1.2 Report
4.1.1.3 Protokoll
4.1.1.4 Lexikoneintrag
4.1.1.5 Handbuch
4.2 Querverweis
4.2.1 Anhang
4.2.2 Glossar

17.12 Lerneinheit „Maße der zentralen Tendenz“ (513), Wissenseinheit „Orientierung“

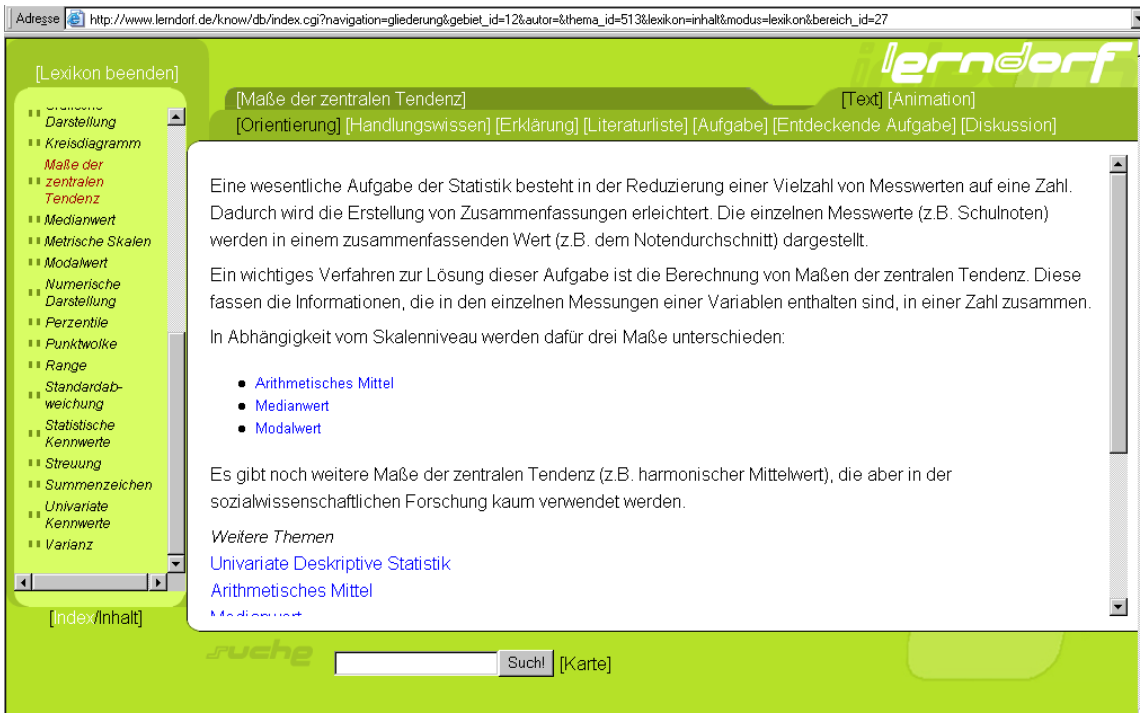

Abbildung 78: Lerneinheit "Maße der zentralen Tendenz" (513); Wissenseinheit "Orientierung".

17.13 Lerneinheit „Arithmetisches Mittel“ (514), Wissenseinheit „Orientierung“

Abbildung 79: Lerneinheit "Arithmetisches Mittel" (515), Wissenseinheit "Orientierung".

17.14 Lerneinheit „Median“ (516), Wissenseinheit „Orientierung“

Abbildung 80: Lerneinheit "Medianwert" (516), Wissenseinheit "Orientierung".

17.15 Lerneinheit „Modus“ (517), Wissenseinheit „Orientierung“

Abbildung 81: Lerneinheit "Modalwert" (517), Wissenseinheit "Orientierung".

17.16 Kreuztabelle der Clusterlösungen

Die folgende Abbildung enthält die Kreuztabelle der Clusterlösungen auf Grundlage der *default*-Definition und der Definition der Substitutionskosten als absoluter Differenz: In der ersten *Zeile* befindet sich die Clusterlösung auf Grundlage der Definition der Substitutionskosten als absolute Differenz; in der ersten *Spalte* befindet sich Clusterlösung auf Grundlage der *default*-Definition der Substitutionskosten.

Anzahl

		cg_subcostabsdiff_ward_28c																											Gesamt
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
cg_subcost default_ward_28c	1	218	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	218
	2	0	0	0	8	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
	3	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
	4	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
	5	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23
	6	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	7	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	3
	10	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	12	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	13	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	4
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	4
	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2
	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	4
	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	10
	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	10	15
	22	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
	23	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19
	24	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	19
	26	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	17
	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35	7	0	42
	28	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
Gesamt		218	14	19	21	12	6	12	23	2	6	9	4	1	1	4	3	2	4	4	2	10	4	2	19	21	35	7	10	475

Abbildung 82: Kreuztabelle der Clusterlösungen auf Grundlage der default-Definition und der Definition der Substitutionskosten als absoluter Differenz.

17.17 Teilkurs: „Statistik - Maße der zentralen Tendenz“

Teillkurs: „Statistik – Maße der zentralen Tendenz“

1	Maße der zentralen Tendenz: OW_T	2642	
2	Maße der zentralen Tendenz: OW_A	3597	
3	Maße der zentralen Tendenz: EW_T	2896	
4	Maße der zentralen Tendenz: HW	2900	
5	Maße der zentralen Tendenz: DIS	3365	513
6	AM: OW_T	2851	
7	AM: OW_A	3185	
8	AM: Bei_T	2863	
9	AM: HW_T	2885	
10	AM: EW_T	2862	
11	AM: MC_TAB	2888	
12	AM: T-F_T	2889	515
13	Summenzeichen: OW_T	2922	
14	Summenzeichen: Bei_T	2928	
15	Summenzeichen: HW_T	2929	
16	Summenzeichen: EW_T	2926	
17	Summenzeichen: MC_T	2948	
18	Summenzeichen: TF_T	2949	615
19	Extremwerte: OW_T	2873	
20	Extremwerte:EW_T	2874	
21	Extremwerte: HW_T	3120	
22	Extremwerte: Bei_T	3121	
23	Extremwerte: TF_T	3123	610
24	Median: OW_T	2790	
25	Median: OW_A	3362	
26	Median: OW_A	2912	
27	Median: Bei_T	2792	
28	Median: Auf_T	2997	
29	Median: HW_T	2886	
30	Median: EW_T	2791	
31	Median: MC_TAB	2891	
32	Median: T-F_TAB	2892	516
33	Modal: OW_T	2848	
34	Modal: OW_A	3363	
35	Modal: Bei_T	2850	
36	Modal: HW_T	2887	
37	Modal: EW_T	2849	
38	Modal: MC_TAB	2899	
39	Modal: T-F_TAB	2897	517
40	Maße der zentralen Tendenz: Auf_T	2999	

17.18 Metadaten: Lerneinheit – Wissensart – Medientyp

Lerneinheit	Wissensart	Medientyp	art_id	thema-id
Maße der zentralen Tendenz	Orientierung	Text	2642	513
	Orientierung	Animation	3597	
	Handlung	Text	2900	
	Erklärung	Text	2896	
	Literatur	Text	3065	
	Aufgabe	Text	2999	
	Entdeck. Aufgabe	Text	2998	
	Diskussion		3365	
Arithmetisches Mittel	Orientierung	Text	2851	515
	Orientierung	Animation	3185	
	Handlung	Text	2885	
	Erklärung	Text	2862	
	Beispiel	Text	2863	
	Literatur	Text	2861	
	True/False	Tabelle	2889	
	Multiple-Choice	Tabelle	2888	
Median	Orientierung	Text	2790	516
	Orientierung	Animation1	3362	
	Orientierung	Animation2	2912	
	Handlung	Text	2886	
	Erklärung	Text	2791	
	Beispiel	Text	2792	
	Literatur	Text	2834	
	True/False	Tabelle	2892	
	Multiple-Choice	Tabelle	2891	
Modalwert	Orientierung	Text	2848	517
	Orientierung	Animation	3363	
	Handlung	Text	2887	
	Erklärung	Text	2849	
	Beispiel	Text	2850	
	Literatur	Text	2860	