EINE STUDIE ZUR EMPIRISCHEN ÜBERPRÜFUNG DER BENUTZERFREUNDLICHKEIT VON TEXTVERARBEITUNGS- UND TABELLENKALKULATIONSPROGRAMMEN

Franz Schiele und Winfried Helge Pelz, awfi Berlin

Zusammenfassung : Ziel der Untersuchung war es, eine Evaluation von Texteditoren und Tabellenkalkulationsprogrammen auf der Grundlage der in der DIN E-66234 Teil 8 genannten Gestaltungsgrundsätze vorzunehmen. Die Eingabesequenzen für die Durchführung eines Anwendungsbeispiels für je 3 Systeme wurden von Beurteilern, die mit der Diskussion um die Benutzerfreundlichkeit vertraut sind, bewertet.

Für die Grundsätze "Aufgabenangemessenheit", "Selbsterklärungsfähigkeit" und "Erlernbarkeit" erhielten wir Einstufungen; die der Aufgabenangemessenheit ließen sich für die Tabellenkalkulationsprogramme anhand des Keystroke-Level-Modells von Card, Moran und Newell vorhersagen.

1 Einleitung : Softwareergonomie und DIN-Gestaltungsgrundsätze

Ein großer Teil der empirischen Forschungsergebnisse der Softwareergonomie resultiert aus einer experimentellen Variation einzelner Systemmerkmale: Untersucht wurden die Auswirkungen unterschiedlicher

- Menüstrukturen (u.a. Perlman 1984);
- Befehlscodierungen (Barnard et.al. 1984, Rosenberg 1983);
- Befehlssprachenstrukturen (Green u. Payne 1984, Carroll 1982);
- Bildschirmstrukturierungen und Codierungen (Benz u. Haubner 1983).

Auswirkungen dieser variierten Systemmerkmale (in der Terminologie der Experimentalpsychologie die sogenannten unabhängigen Variablen) wurden u.a. in folgenden Meßgrößen bestimmt:

- Zeit, z.B. zum Suchen einer Option in einem hierarchischen Menüsystem (Perlman 1984) oder zum Ausfüllen einer Dateneingabemaske (Benz u. Haubner 1983);
- Fehlerhäufigkeit und Zeit, die zur Korrektur von Fehlern beansprucht wird (Williges et al. 1984);
- Urteile und Bewertungen der Benutzer in den verschiedensten Dimensionen; subjektive Ratingverfahren, Interviews etc., z.B. de Bachtin (1984) oder auch Tyhan (1984);
- Lernleistungen, in der Art klassischer Gedächtnispsychologie (Green u. Payne 1984);
- Häufigkeit der Benutzung von Befehlen und von Helpfunktionen (Williges et al. 1984);
- Handlungs- bzw. Planungstrukturanalysen, d.h. welche verfügbaren Kommandos werden in welchen Aufgabensituationen benutzt (Riley u. O´Malley 1984), zu erheben mittels der Playbackmethode (Neal u. Simons 1984).

Diese Meßgrößen (in der Experimentalpsychologie die sogenannten abhängigen Variablen) werden herangezogen, um Konzepte wie "Ease of use" und "Ease of learning" zu erfassen.

Unabhängige Variablen z.B.		Abhängige Variablen
Menüstrukturl Menüstruktur2	⟶	Auswahlzeiten

Empirische Ergebnisse dieser Art sind für den Systemdesign- und verbesserungsprozeß durchaus anwendbar. Um aber als praktikable Grundlage für die softwareergonomische Evaluation ganzer Systeme dienen zu können, müssen die Einzelbefunde in ein differenziertes Kriterienraster integriert werden.

Für den Anwender ist gegenwärtig bei der Auswahl unter bestehenden Softwaresystemen deren vergleichende ergonomische Bewertung von Bedeutung.

Roberts (1980) und Roberts u. Moran (1983) haben ein Bewertungsverfahren anhand verschiedener Texteditoren vorgeschlagen. Sie haben folgendes erhoben:
Die Zeit, und Fehlerkorrekturzeit, die "Experten" auf diesen Systemen für vorgegebene Benchmarkaufgaben benötigen, den Lernerfolg von "Anfängern" und eine Bewertung der Funktionalität der Systeme, d.h. welches Aufgabenspektrum mit ihnen durchfürbar ist, s.a. Borenstein (1985).

Unabhängige Variablen		Abhängige Variablen
System A		Zeit
System B	→	Fehlerzeit
System C		Erlernbarkeit
etc.		Funktionalität

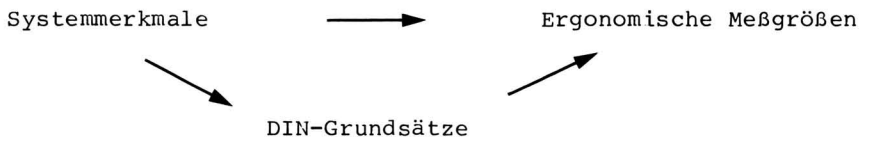
Die DIN 66234 Teil 8 "Bildschirmarbeitsplatze - Grundsätze der Dialoggestaltung" (Entwurf Dez. 84) legt ergonomische Anforderungen an Dialogsysteme fest: es "sollen die Eigenschaften des Systems an die psychischen Eigenschaften der damit arbeitenden Menschen angepaßt werden." Die fünf Gestaltungsgrundsätze sind positiv definierte Eigenschaften des "Zusammenwirkens von Mensch und System", deren technische Realisierung bewußt offengelassen wird.

An ein Prüfverfahren sind mindestens folgende Anforderungen zu stellen:

1. es muß arbeitswissenschaftlich fundiert sein,
2. zu Aussagen führen, die im Anwendungsfeld praxisnah sind und
3. es muß ökonomisch durchführbar sein.

Die Bewertung eines Systems anhand solcher Grundsätze setzt deren Meßbarkeit voraus.

Eine arbeitswissenschaftliche Fundierung kann an den oben angeführten Einzeluntersuchungen nicht vorbeigehen. Es erscheint daher sinnvoll, zu versuchen, einzelne technische Systemmerkmale und deren ergonomische Auswirkungen den DIN-Grundsätzen zuzuordnen:



2 Untersuchungsziel

In der vorliegenden Untersuchung wurde versucht, gemäß den in den DIN-Entwürfen diskutierten Grundsätzen eine Evaluierung von Texteditoren (TE) und Tabellenkalkulationsprogrammen (TK) vorzunehmen, um dabei sowohl Hinweise für die Konkretisierung der Grundsätze (in welchen Systembereichen finden sich Eigenschaften, die sich den verschiedenen Kriterien zuordnen lassen), als auch für geeignete Verfahren ihrer Überprüfung zu erhalten.

3 Methode

Jeweils drei kommerziell erhältliche TE und TK wurden von je 10 Experten beurteilt, die alle über Erfahrungen mit Systemen der jeweiligen Kategorie verfügten und mit der softwareergonomischen Diskussion zur Benutzerfreundlichkeit von Dialogsystemen vertraut waren.

Im Unterschied zu durchschnittlichen "echten" Benutzern konnte von diesem Personenkreis erwartet werden, daß er bei der Systembeurteilung auf allgemeinere systemunabhängige Konzepte rekurriert, in denen die ergonomische Relevanz einzelner Systemmerkmale mitberücksichtigt wird (s.a. Hammond et al. 1984).

Vor dem Systemvergleich wurde den Beurteilern eine Übersicht der in den DIN-Entwürfen diskutierten Gestaltungsgrundsätze vorgelegt, deren Kurzdefinitionen sie - soweit ihnen das notwendig oder sinnvoll erschien - ergänzen bzw. verändern sollten.

3.1 Beurteilung der Softwaresysteme

Um eine möglichst praxisrelevante Beurteilung zu erhalten, wurde ein konstruiertes Anwendungsbeispiel gezeigt, das aus einer Folge von Einzelaufgaben bestand. Zu bewerten war die Realisierung dieser Aufgaben in den 3 Systemen anhand einer schriftlich-symbolisch dokumentierten Vorlage der jeweils erforderlichen Eingabesequenzen.

3.2 Konstruktion des Anwendungsbeispiels

3.2.1 Aufgabeninhalte

Als Anwendungsbeispiel wählten wir die Durchführung verschiedener Korrekturen an einer vorgegebenen Textseite bzw. Tabelle. Hierbei wurden solche Prozeduren herangezogen, die bei der Bearbeitung von Dokumenten mit TE bzw. TK besonders häufig vorkommen: u.a. Cursorpositionierung, Einfügen, Löschen, Kopieren, Versetzen, "Suchen und Austauschen" von Textteilen etc. bzw. Eingeben von Text-, Zahlen- und Formeleinträgen, Formatieren, Kopieren, Versetzen, Löschen von Zelleneinträgen, Zeilen oder Spalten (sog. "core editing tasks", Roberts 1980).

3.2.2 Aufgabenumfang

Die einzelnen Aufgaben umfassen häufig die Benutzung mehrerer Systemfunktionen, denn die Objekte, auf die sich Systemfunktionen beziehen, sind oft nicht kongruent mit den semantischen Strukturelementen eines Dokumentes (sog. "task attributes" Smolensky et al. 1984): Um beispielsweise das letzte Wort eines Satzes zu löschen, kann es erforderlich sein, nach der Funktion "Wort löschen" den Punkt am Satzende neu einzufügen, wenn "Wort" systemintern als eine durch Leerzeichen begrenzte Zeichenfolge definiert ist und somit der Punkt mitgelöscht wird.

Ebenso (und eng damit zusammenhängend) unterscheidet sich die systemspezifische Aufgabenstruktur oft von der Struktur der Aufgabe, wie sie mit herkömmlichen Mitteln ausgeführt würde: Um z.B. die Werte einer Rubrik innerhalb einer Tabelle zu berechnen, wird die entsprechende Formel in eine Zelle der Rubrik eingetragen und dann der Inhalt dieser Zelle (d.h. diese Formel) in die übrigen Zellen kopiert. Mit anderen Worten: Aufgaben müssen erst in eine Folge von Systemfunktionen "übersetzt" werden. Selten entsprechen einzelne Systemfunktionen unmittelbar den Aufgaben des Benutzers (so stellt das Kopieren eines Zelleninhalts für sich genommen keine sinnvolle Aufgabe dar).

Die Komplexität oder auch Einfachheit dieser "Übersetzung" von systemunabhängig formulierten Aufgaben in die systemspezifischen Operationen, von Moran (1983) als "external-internal task mapping" beschrieben, trägt entscheidend zur Benutzerfreundlichkeit eines Systems bei. Deren Erfassung setzt die Durchführung von Aufgaben voraus, die in Begriffen des externen Aufgabenbereichs (external task space) definiert sind.

3.2.3 Aufgabenspezifizierung: Festlegung der Eingabesequenz

Bei der Konstruktion von Aufgaben, insbesondere solchen, die mehrere Operationen umfassen, ergibt sich häufig das Problem, daß der Arbeitsweg durch das gewünschte Ergebnis nicht eindeutig determiniert ist: unterschiedliche Eingabesequenzen können zum gleichen Ergebnis führen. Dabei kann es sich um unterschiedlich günstige Varianten oder um Freiheitsgrade (gleichgünstige Varianten) handeln.

Die den Beurteilern vorgegebenen Sequenzen wurden so gewählt, daß sie die systemspezifisch optimale Lösung für die gegebene Aufgabe darstellten (optimal im Sinne der Anwendung der jeweils mächtigsten Funktionen bzw. möglichst kurzer Eingabesequenzen). Soweit die Systeme unterschiedlich mächtige Funktionen besitzen, wurden diese auch in den vorgegebenen Eingabesequenzen repräsentiert und nicht etwa einer systemübergreifend möglichst einheitlichen Aufgabenbearbeitung "geopfert": Primär sollte beurteilt werden, wie die selbe Aufgabe möglichst effektiv ausgeführt wird, und nicht etwa, wie die gleiche Funktion bei den verschiedenen Systemen realisiert ist.

Sofern die Systeme jedoch mehrere gleichgünstige Varianten boten, wurde bei der Festlegung der Eingabesequenzen auch eine möglichst ähnliche Aufgabenbearbeitung angestrebt.

4 Untersuchungsablauf

Zur Identifizierung der Aufgaben lag den Beurteilern je ein Ausdruck des Dokuments, auf das sich die erforderlichen Eingabesequenzen bezogen, im Ausgangszustand (mit handschriftlichen Korrekturanweisungen) und im Endzustand (nach Ausführung der Aufgaben) vor. Jede Aufgabe war außerdem allgemein, d.h. ohne Bezug auf Eigenschaften der Systeme, in termini des gewünschten Ergebnisses beschrieben, z.B. "Für die Rubriken X und Y sollen die Summen berechnet werden".

Wenn die Aufgabe mehrere Dialogschritte umfaßte, waren diese zusätzlich beschrieben; diese Beschreibungen rekurrierten bereits auf Objekte und Funktionen des jeweiligen Systemtyps (TE oder TK), waren aber noch systemübergreifend formuliert: z.B. "Zellen-Eintrag eingeben (Formel)", "Zellen-Eintrag kopieren".

Die systemspezifische Beurteilungsgrundlage bildete schließlich die schriftliche Darstellung der Eingabesequenzen, die zur Ausführung der Aufgabe bzw. der einzelnen Dialogschritte erforderlich waren. Diesen Vorlagen waren schriftliche Erläuterungen zur Bedeutung der dargestellten Tastenanschläge (ihre Funktionen) beigefügt.

Die Aufgabe der Beurteiler bestand darin, sich jeweils für eine Aufgabe die Abfolge der Tastenanschläge bzw. ihre Funktionen zu vergegenwärtigen und diesen Prozeß der mentalen Repräsentation dem Versuchsleiter durch "begleitendes Sprechen" mitzuteilen.

Nachdem sie sich eine Aufgabe bei allen 3 Systemen angesehen hatten, sollten sie diese systemspezifischen Realisierungen hinsichtlich der oben erwähnten DIN-Grundsätze auf einer Rating-Skala von 8 (sehr gut) bis 1 (sehr schlecht) bewerten. Anmerkungen und Urteilbegründungen wurden protokolliert.

5 Ergebnisse

Von den vorgelegten Gestaltungsgrundsätzen Aufgabenangemessenheit, Selbsterklärungsfähigkeit, Erlernbarkeit, Steuerbarkeit, Verläßlichkeit, Fehlertoleranz/-transparenz und Flexibilität wurden lediglich

- Aufgabenangemessenheit
- Selbsterklärungsfähigkeit und
- Erlernbarkeit

bewertet.

Die übrigen Grundsätze sahen fast alle Beurteiler auf der Grundlage der zur Verfügung stehenden Informationen (d.h. bei der Vorlage einer festgelegten, fehlerfreien Aufgabenabarbeitungssequenz) als nicht bewertbar an.

Dabei wurden Selbsterklärungsfähigkeit und Erlernbarkeit nur bedingt als zwei unterschiedliche Gestaltungsgrundsätze betrachtet wurden, denn ihre Korrelation beträgt TE r = 0.82 bzw. TK r = 0.85 .

Die Experten wiesen wiederholt darauf hin, daß sie zur Beurteilung der übrigen Gestaltungsgrundsätze zusätzliche Informationen benötigen würden, z.B. die Vorgabe unterschiedlicher Arbeitswegvarianten zur Beurteilung der Steuerbarkeit, oder Angaben über das Systemverhalten bei Fehlbedienung zur Beurteilung von Fehlertoleranz und -transparenz.

In Folgeuntersuchungen werden wir die Informationsgrundlage der Beurteiler systematisch erweitern.

Die Bewertungen fallen bei demselben System über die verschiedenen Aufgaben hinweg höchst unterschiedlich aus. Es lassen sich also auf diese Weise für aufgabenspezifische Funktionsbereiche Polaritätsprofile erstellen, und damit Schwachstellen bzw. Stärken aufweisen.

Da die Beurteiler ihre Einstufung zur Aufgabenangemessenheit häufig mit einer Aufwandsabschätzung begründeten, interessierte uns, inwieweit diese mit der Länge der geforderten Eingabesequenzen zusammenhängen.

Eine Analyse der Bewertungen mit einem vereinfachten Keystroke-Level-Modell nach Card, Moran und Newell (1983), bei dem die Anzahl der Tastenanschläge (keystrokes = k) und die Anzahl der mentalen Operationen (mentals = m) für die

einzelnen Aufgaben ausgezählt wurden ergibt mit den Bewertungen folgende Korrelationen :

keystrokes/Aufgabenangemessenheit

TE $r = -0,35$
TK $r = -0,56$

mentals/Aufgabenangessenheit

TE $r = -0,40$
TK $r = -0,72$

Nimmt man keystrokes und mentals als Prädiktoren für die Bewertung der Aufgabenangemessenheit, erhält man multiple Korrelationen von

TE $R_{3.12} = 0,42$
TK $R_{3.12} = 0,72$

wobei die Standardpartialregressionskoeffizienten für die keystrokes

TE $b_1 = -0,18$
TK $b_1 = -0,12$

und für die mentals

TE $b_2 = -0,29$
TK $b_2 = -0,63$

betragen. Bei den TKs entspricht das in etwa der Gewichtung, die Card, Moran und Newell den Einheiten als Zeitmultiplikator gegeben haben.

Es erscheint somit möglich Bewertungen der Aufgabenangemessenheit von Tabellenkalkulationsprogrammen mit hinreichender Genauigkeit aufgrund eines analytischen Modells vorherzusagen.

6 Literaturverzeichnis

Bachtin, de, O.: It is what it´s used for - job perception and system evaluation. Conference Papers INTERACT ´84, London, Vol. 2, 199-202.

Barnard, P.; MacLean, A.; Hammond, N.: User representations of ordered sequences of command operations. Conference Papers INTERACT ´84, London, Vol. 1, 434-438.

Benz, C.; Haubner, P.: Codierungswirksamkeit bei Informationsdarstellungen in Bildschirmmasken. In: Balzert, H. (Hrsg) Software-Ergonomie. Stuttgart 1983.

Borenstein, N. S.: The evaluation of text editors: A critical review of the Roberts and Moran methodology based on new experiments. Proc. CHI´ 85 Human Factors in Computing Systems (San Francisco, April 14-18, 1985), ACM, New York, 99-105.

Card, S. K.; Moran, T. P.; Newell, A.: The Psychology of Human-Computer Interaction. Hillsdale,N.J. 1983.

Carroll, J. M.: Learning using and designing command paradigms. Human Learning, 1, (1982) 31-62.

DIN-66234 Teil 8 , Bildschirmarbeitsplätze: Grundsätze der Dialoggestaltung. Dez 1984

Green, T. R. G.; Payne, S. J.: Organization and learnability in computer languages. International Journal of Man-Machine Studies, 21, (1984) 7-18.

Hammond, N. et al.: Evaluating the interface of a document processor: a comparison of expert judgement and user observation. Conference Papers INTERACT ´84, London, Vol. 2, 135-139.

Moran, T.P.: Getting into a system: external-internal task mapping analysis. Proc. CHI ´83 Human Factors In Computing Systems, New York: ACM (1983) 45-49.

Neal, A. S.; Simons, R. M.: Playback: A method for evaluating the usability of software and its documentation. IBM Systems Journal, 23, (1984) 82-96.

Perlman, G.: Making the right choices with menus. Conference Papers INTERACT ´84, Vol. 1, 291-295.

Riley, M.; O´Malley, C.: Planning nets: A framework for analyzing user-computer interactions. Conference Papers INTERACT ´84, London, Vol. 1, 86-91.

Roberts, T. L.: Evaluation of computer text editors. Ph.D. dissertation, Departement of Computer Science, Stanford University, Stanford, Calif., (1980).

Roberts, T. L:; Moran, T. P.: The evaluation of text editors: Methodology and empirical results. Communications of the ACM, 26, (1983) 265-283.

Rosenberg, J.: A featural approach to command names. Proc. CHI ´83 Human Factors in Computing Systems, New York, ACM (1983) 116-119.

Smolensky, P.; Monty, M. L.; Conway, E.: Formalizing task descriptions for command specification and documentation. Conference Papers INTERACT ´84, London, Vol. 1, 16-22.

Tynan, P. D.: Randomly sampled self-report method for collecting field data on human-computer interaction. Conference Papers INTERACT ´84, London, Vol. 2, 209-212.

Williges, R. C. et al.: Providing online assistence to inexperienced computer users. Conference Papers INTERACT ´84, London, Vol. 2, 113-117.

Franz Schiele

Winfried Helge Pelz

Arbeitswissenschaftliches Forschungsinstitut, awfi GmbH

Bayreuther Str. 3

D - 1000 Berlin 30