ヒューマンインタフェース 63－8
（1995．11．2）

対話履歴によるユーザビリティ評価

旭敏之，岡田英彦，井関治

ＮＥＣ関西Ｃ＆Ｃ研究所

対話履歴をベースにした使いやすさ評価方式の研究を進めるためには，理論的基盤として「対話履歴分析で何がどこまで評価できるか」を明確にする必要がある．ファクシミリ等のパネル型ＵＩを対象とした使いやすさ評価システムＵＩテスタを用いて，被験者30人による大量の対話履歴データを分析するとともに，同じ対象をヒューリスティック評価手法で評価した．両者の結果を比較した結果，対話履歴ベースの評価でも２／３程度の問題点が抽出できること，用語やエラー防止に関するものは抽出されやすいが，一貫性や作業効率に関する問題点は抽出されにくいことなどが明らかになった．

Usability Evaluation Based on User Interaction Log Analysis

Toshiyuki Asahi,　Hidehiko Okada,　Osamu Iseki

Kansai C&C Research Laboratories, NEC Corporation
1-4-24, Shiromi, Chuo-ku, Osaka 540, Japan
(asahi@obp.cl.nec.co.jp)

Fundamental theories providing what extent of usability can be evaluated with interaction log analysis techniques are required to promote the research on developing the automatic usability testing system. In order to record and analyze a large set of interaction log data with UI-tester, an experimental usability testing for a facsimile machine was conducted with the cooperation of thirty subjects. The same facsimile was also tested with Heuristic Evaluation method. By comparing those two set of result, following points are clarified; Two third of usability problems can be extracted with log data based evaluation. Problems concerning terms or visibility and error prevention are relatively easy to extract even with log data based evaluation. On the contrary, problems about consistency and efficiency are hard to find out.

1．はじめに

ユーザビリティ・テスティングを製品開発工程に組み込むためには，相応のリソースが必要であり，特にヒューマンファクタ専門家の数が少ない日本メーカでは対応が難しい．筆者らは，工数を削減し均質な評価結果を得るため，使いやすさの自動評価システム「ＵＩテスタ」の開発を進めてきた．

ＵＩテスタでは，ユーザと評価対象システムとの対話履歴（"ユーザの入力"とそれに対する"システムの応答"を時系列でコード化したもの）を分析し，ユーザビリティ上の問題点を抽出する．一般に，対話履歴はユーザビリティ評価において貴重なデータとされているが[1]，データにユーザの操作意図に関する情報が含まれないため，完全な評価が困難なことも十分予想される．自動評価システムの研究を進めるためには「対話履歴をベースにして何がどの程度評価できるか」を明確にしておくことが不可欠である．

今回，被験者30人の協力を得て大量の対話履歴データを収集しＵＩテスタで分析した．さらに同じ評価対象（ファクシミリ）をヒューリスティック評価方式で評価し，両者の結果を比較することで，対話履歴ベースによるユーザビリティ評価の特性を明確にした．以下，これらの検討結果を報告する．

2．対話履歴ベース評価方式

2.1 従来技術

使いやすさ評価のためのツール／システムは，ユーザインタフェース（UI）そのものを何らかの記述形式に変換し分析するものと，ユーザ対話履歴をベースにするものとに大別することができる．前者では，一貫性や対話部品の配置妥当性など，その記述形式で明らかになる側面に絞って検証するのに対し[2][3]，後者では設計者が予想できなかった問題点を発見できることが特徴である．後者はさらに次のように分類することができる．

1) プロトコル解析やモニタリング評価手法を補助するために，対話履歴を記録するもの．
2) 対話履歴の分析機能を提供し，使いやすさ評価作業をサポートするもの．

1) は操作状況を録画したテープの読み取り（編集，検索，アノテーションなど）を補助するための機能が主であり，特に分析機能は提供されない[4]．2) では大量に得られる対割履歴データを評価の容易な形式に変換したり，そこから特徴的なパタンを抽出したりすることが特徴である．繰り返しパタンを発見する手法[5]，状態遷移パタンを分析する手法[6]，標準の対話手順との差異を検出する手法[7, 8]などが報告されている．2) の方式は，自動記録可能な対話履歴だけを用いて分析できる点が長所であるが，提案された手法が実際のユーザビリティ・テスティングにおいてどの程度有効であるか，あるいはユーザビリティの問題点をどれだけ網羅して抽出できるかに関しては未だ十分検討されていない．

2.2 ＵＩテスタ

2.2.1 基本コンセプト

ＵＩテスタは，ユーザ対話履歴と標準の対話手順とを比較しその相違部分を抽出することで，使いやすさの問題点を発見する[8−10]．対話履歴にユーザのメンタルモデルが，標準対話手順にはＵＩ構造のそれぞれ一部が反映されていると仮定し，この両者の比較分析を行うことが基本的なコンセプトである．

まず，評価の対象をファクシミリなどのパネル型ＵＩに絞って分析機能を構築した．パネル型ＵＩでは標準の対話手順が１本のルートで表現され，そこから逸

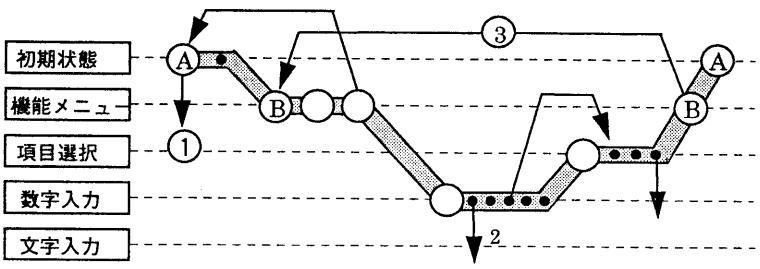

図１　対話構造ダイアグラム[8-10]

脱したものは誤操作であると判定できるからである．これを評価者に明示的かつ直観的に提示するため，標準の対話手順を太い帯で，そこから逸脱したユーザ対話を細いアークで表現する対話構造ダイアグラムを考案した．概略を図１に示す．

2.2.2 使いやすさ評価機能

ＵＩテスタの主な対話履歴分析機能としては，共通誤対話分析機能と共通操作時間分析機能がある．以下，その概略を示す．

共通誤対話分析機能 [11]

一般に誤対話はユーザビリティの問題点を示唆する重要な指標であるが，実際の誤対話パタンの中には個人差によるものや偶然に生じたものなどが含まれており，そのすべてが直接問題点を示しているとは限らない．ＵＩテスタでは，複数の被験者間で共通的に見られた誤対話パタンを選択的に抽出する機能を有している．共通性を判定する際には「視点」「厳密度」「共通度」の３つのパラメータが操作できるようになっており，評価者がインタラクティブに誤対話パタンを絞り込んでいくことができる．抽出された誤対話は対話構造ダイアグラム上に表示される．

共通操作時間分析機能 [12]

ユーザの操作時間は製品のユーザビリティを検討する上で重要な指標である．特に良く統制された実験環境下では信頼性が高く定量的な分析が可能となる．ＵＩテスタでは，対話履歴を記録する際には時刻データを同時に記録し，これを用いた操作時間分析機能を提供している．ただし，上記と同様に操作時間からユーザビリティ上の問題点を確定的に議論することは難しいので，やはり複数ユーザの操作時間データから，共通的な傾向（平均値と標準偏差値）を計算し，これを評価者に提示する．

３．対話履歴収集実験

3.1 実験Ｉ：製品開発への適用 [13]

ＵＩテスタが実際の製品評価の場面で有効であるかどうかを検証するため，実際に開発段階にあったファクシミリ製品に対し，ユーザビリティ評価の未経験者が評価テストを実施した．被験者 18 人に対してそれぞれ３つのタスクを与え，得られた対話履歴を上記評価機能を用いて分析した．９個の改善点が指摘され（内容は表２の中で示す），そのうちの５箇所は製品出荷前に何らかの改善が施された．

3.2 実験ＩＩ：対話履歴分析の有効性検証

3.1 の実験は，ＵＩテスタが実際の製品開発過程で活用できるか否かを検証したものである．したがって，抽出された問題点は「ユーザビリティ評価の未経験者」が容易に指摘できたものに限られている．はじめにも述べたように，本研究の目的は「対話履歴分析でユーザビリティの何がどの程度評価できるか」を明らかにすることにある．すなわち，対話履歴でどの範囲の製品問題点を発見することができるかを見極める必要がある．この分析の精度を上げるためには，さらに大量の対話履歴データが必要である．実験Ⅱはこの目的のため実施された．概要を以下に示す．

- 被験者：30 名（初心者 15 名，中級者 15 名）
- 評価対象：実験Ｉと同じファクシミリ製品．
- タスク：実験Ｉで用いたものと同じタスク．ただし，各被験者のテスト時間を１時間に区切ったため，被験者によっては実行できなかったタスクも存在する．

４．評価能力の検証

対話履歴ベース評価の有効性を検証するためには，基準となる評価結果を設定する必要がある．本研究では，すでに利用実績が多く実践的な技法として評価が高いヒューリスティック評価手法 [14, 15] を用いる．

4.1 ヒューリスティック評価

プロトコル解析やユーザモニタリング方式の評価方式を簡略化し，テストに要するコストを削減するために考案された手法である．ユーザ参加を前提とせず，複数の専門家が評価対象を吟味し，集中的に問題点を抽出する点が特徴である．評価の指針として表２に挙

表２ 評価に用いるヒューリスティック [14][15]
現在は新しいセットが提案されているが [16]，本研究では実用実績が多いオリジナル版を採用した．

①シンプルで自然な対話方式
②ユーザの言葉を使用する
③ユーザの記憶負担を最小にする
④一貫性を保つ
⑤ユーザにフィードバックする
⑥出口を明示的に提供する
⑦ショートカットを提供する
⑧良いエラーメッセージを提供する
⑨エラーを未然に防ぐ
⑩ヘルプとドキュメントを提供する

表3 抽出された問題点

ヒューリスティクス欄の番号は，表2の各ヒューリスティクを示しており，各評価者が申告したものである．

項目	ユーザビリティ問題点	評価者	タスク	ヒューリスティクス	対応履歴 I	対応履歴 II
H1	[発信元情報]というラベルがわかりにくい．[自局]と[発信元]の区別がわかりにくい．	ABCD	2	②⑨	○	○
H2	電話番号と発信元情報は同じ画面で登録できた方が良い．	ABD	2	①⑦		
H3	[全／半]キーなどで，操作可能な状態とそうでない状態を明示的に区別すべき．	ABD	3	①③⑤⑨	○	○
H4	宛先登録でボタンイメージの中に文字を入力していることが，利用者側に伝わらない．	ABD	3	①		
H5	[ワンタッチ]キーと[ワンタッチプログラム]キーが紛らわしい．	ACD	3	②⑨	○	○
H6	[海外［○×］]など，操作不可能な箇所に直接タッチしたくなる．	ABC	3	⑨①	○	○
H7	同機能に[削除]と[1文字削除]の2通りのキーラベルが使われている	ABD	1	④①②		
H8	漢字変換候補確定のために[完了]キーを押下するのは不自然．	BCD	2	①②⑨	○	○
H9	[機能登録][宛先登録][初期登録]や[現在時刻]においてボタンによる	AB	1	①⑨		○
H10	現在時刻入力画面で，入力手順のガイダンスが欲しい．	AB	1	①⑩		○
H11	[確定]キーを[変換]キーの隣に配置すべき．[確定]キーを押すのに抵抗がある．	AD	2	①⑨	○	○
H12	矢印キーがどの表示オブジェクトに対し有効かが不明確．[海外]や[通信モード]で矢印ｷｰが使えないのが不自然．	AD	3	①④		○
H13	自局電話番号登録やワンタッチキー登録では[取消]キーが無い．	AD	2	④⑦		
H14	[クリア]の意味が[現在時刻登録]と[発信元登録]との場面で異なる．	BD	1	④②		
H15	[現在時刻]は[現在日時]というラベルの方が正確．	B	1	①⑨		○
H16	漢字変換できなかったときのフィードバックがない．	C	2	⑤⑨	○	○
H17	時刻登録が正常に行われたかどうかのフィードバックが欲しい．	C	1	⑤⑨		○
H18	[自局電話番号]と[自局市外局番]キーが紛らわしい．	A	2	⑨①	○	○
H19	未登録の[ワンタッチ]キーは未登録であることを明示すべき．	A	3	①⑨		
H20	カナ入力モードの時はキーラベルもカナ表記になるべきである．	A	3	①②		
H21	表示データの区切りが狭すぎて見づらい．	B	3	①⑨		
H22	カナ記号の（ー）の入力方法がわからない．	C	3	①③	○	○
H23	[プログラム][自局短ダイアル]の言葉の意味がわかりにくい．	C	3	②		○
H24	[ワンタッチ削除]を押した後に確認が無い．	D	3	⑨⑤		
H25	ポーズ記号（-）の入力方法がわからない．		2		○	

げるようなヒューリスティクスが規定されており，各評価者はこれを問題点抽出の判断基準とすると共に，抽出した各問題点にどのヒューリスティクスに関連するかを明示することが求められる．テスト工数短縮の効果は大きいことに加え，問題点抽出能力に関してもフォーマルな評価方式に較べて遜色無いことが報告されている[15]．

4.2 評価結果

4人のHI研究者が参加し，3で対象としたものと同じ機種を評価した．ヒューリスティック評価では本来はタスクを限定しないが，今回は比較のため実験と同じタスクだけを試行し問題点を洗い出した．表2にヒューリスティック評価で抽出された問題点を列挙する[注]．表2において，1～24はヒューリスティック評価で抽出された問題点である．各問題点を指摘した評価者（ABCD）とその問題点に関連するヒューリスティクスも付記する．さらに，これらの問題点のう

表4 ヒューリスティックと関連問題点

ヒューリスティック	A	B	C
シンプルで自然な対話方式	17	11	65
ユーザの言葉を使用する	7	4	57
ユーザの記憶負担を最小にする	3	3	—
一貫性を保つ	4	1	—
ユーザにフィードバックする	4	3	—
出口を明示的に提供する	0	0	—
ショートカットを提供する	2	0	—
良いエラーメッセージを提供する	0	0	—
エラーを未然に防ぐ	14	11	79
ヘルプとドキュメントを提供する	0	0	—

A：関連する問題点
B：Aのうち，実験Ⅰ，Ⅱで抽出されたもの
C：B/A×100 (%)　(ただし，Aが5以上の項)

表5 操作時間分析により抽出された問題点

問題	視点	タスク	問題点
T1	I	1	初期画面で操作にまよう
T2	I	1	H10
T3	I	2	かな入力にとまどう
T4	II	2	H1
T5	II	2	H11
T6	II	3	H3
T7	II	3	H12

ち実験Ⅰで検出されたものを「Ⅰ」の欄の○で示し，さらに実験Ⅱで得られたデータを分析し，各問題点を示唆するパタンが認められた（＝問題箇所で誤対話パタンが観測された）ものには「Ⅱ」の欄の○で示す．後者のケースは，条件の厳しいⅠの試行実験では問題点として指摘されなかったが，対話履歴の中に「問題点抽出のための情報が含まれているもの」である．すなわち，基本的には分析方式を工夫することで抽出可能な問題点を示していると解釈できる．

表3から明らかなように，今回検出された総数25の問題点の内訳は，

- ヒューリスティック評価で抽出でき，対話履歴でも確実に抽出できるもの … 9
- ヒューリスティック評価で抽出でき，特徴的なパタンが観察できるもの … 6
- ヒューリスティック評価で抽出でき，対話履歴では抽出できないもの … 9
- 対話履歴分析で抽出でき，ヒューリスティック評価で見落としたもの … 1

となる．

さらに評価者が指摘したヒューリスティクスに着目し，どういった種類の問題点が対話履歴から発見しやすい（にくい）かを検討する．表4に各ヒューリスティックに関連する問題点と，そのうち実験Ⅰ，Ⅱのそれぞれで発見された問題点の数を集計する．

4.3 操作時間による評価

ＵＩテスタでは，2.2.2に示したように各操作ごとに複数ユーザの操作時間の平均値を得ることができる．今回の実験では，視点Ⅰ：「正解の操作を行うのに多大の時間を要している」，視点Ⅱ：「短い操作時間で誤操作を行っている」という2種類の現象が問題点の候補を表していると考え，それに該当する箇所を抽出した．どの時間を境に「多大の操作時間」とするかを明確に規定することは難しいが，今回は10秒を一応の目安とした．その結果Ｔ1～Ｔ7の7つの問題点が抽出された（詳細は文献[12]参照）．表5に示すようにそのうち5項目はヒューリスティック評価と実験Ⅱの双方で検出されたものである．操作時間分析で新たに抽出されたＴ1とＴ3は，ユーザが新しい操作環境に遭遇して時間が長くかかったものであり，対話パタンからは発見できなかったものである．ただし，未知の場面に遭遇して操作に余分の時間を要することは当然でもあり，ＵＩの問題点として取り上げるべきものであるかどうかは疑問である．以上の結果を勘案すると，操作時間分析は問題の性質をより詳細に分析したい場合や，対話パタンによる問題点抽出数を向上させたい場合などのために補助的に用いるべきであると考えられる．

5．考察

本研究では，モニタ実験による大量データをベースに「対話履歴でどこまでユーザビリティ評価が可能か」という命題に対して実験的に検討を加えた．以下，得られた結果から対話履歴分析の特性を考察する．

・ヒューリスティック評価を基準とした場合，約2/3

注) ただしここでいう「問題点」とは，あくまで評価者の一方的な意見をまとめたものであり，製品として改善すべきものかどうかはこの段階では不明である．より正確には「問題点の候補」とすべきであるが，冗長であるため単に「問題点」と称する．

（16/25）の問題点は対話履歴から発見可能であり，約1/3（9/25）は現行テスタの機能でかつ初心者でも確実に抽出できる．

・ヒューリスティクスとの関連では，以下の点が明らかになった．

－「エラーを未然に防ぐ」に関する問題点の発見可能性が高い（79%）．これは，この種の問題が直接的にユーザの誤操作となって表面化しやすいためであると考えられる．

－「ユーザの言葉を使用する」は，ＵＩで使われている用語がわからないケースを示しており，特に今回のようなパネル型インタフェースで発生しやすい問題を示している．用語がわからない結果誤ったボタンを押下しやすいため，対話履歴で発見できるケースが多くなっている（57%）．

－その一方で，ケースは少ないが一貫性の問題は発見可能性が低い．一般的にこの種の問題はヒューリスティック評価では発見されやすいが，必ずしも誤対話となって表面化するとは限らない．したがって，対話履歴では発見しにくいものとなっている．

－「ショートカットの提供」も発見可能性が低いことが予想される．これは，「多少操作が不便であっても誤対話までは誘発しない」ようなケースでは対話履歴から発見することが難しいためである．

・操作時間分析は，問題の性質をより詳細に分析したい場合や，対話パタンによる問題点抽出の精度を上げたい場合などのために補助的に用いるべきである．

以上のように，大量のデータを分析することで，対話履歴分析による使いやすさ評価の傾向を掴むことができた．もちろん，今回の研究では評価対象や分析ツールが限定されており，一般的な議論に展開するには不十分である．また，GUIなどパネル型とは特性の違うUIでも同様の分析を行う必要がある．今後の研究課題としたい．

参考文献

[1] Nielsen, J. : Usability Engineering, Academic Press, Inc., 1993.

[2] Sears, A. : Layout Appropriateness: A metric for widget-level user interface layout evaluation, Technical Report CAR-TR-603, CS-TR-2838, University of Maryland Computer Science Department, 1992.

[3] 神場知成，橋本治：マルチビューモデルに基づくユーザインタフェース設計ツールU-Face, 情報処理学会論文誌, Vol. 34, No. 1, pp. 167-176, 1993.

[4] Weiler, P., Cordes, R., Hammontree, M., Hoiem, D., Thompson, M.: Software for the Usability Lab: A Sampling of Current Tools, Human Factors in Computing Systems, INTER-CHI '93 Proceedings, pp. 57-60, 1993.

[5] Siochi, A., Ehrich, R.: Computer Analysis of User Interface Based on Repetition in Transcripts of User Sessions", ACM Transactions on Information systems, Vol. 9, No. 4, pp.309-335, 1991.

[6] Guzdial, M.: Deriving Software Usage Patterns from Log Files", Georgia Institute of Technology Technical Report, 1994.

[7] Kishi, N.: SimUI: Graphical User Interface Evaluation Using Playback", Proceedings of the 16th Annual International Computer Software and Applications Conference, pp. 121-127, 1992.

[8] 旭敏之，井関治，岡田英彦：「使いやすさ」の計測：ＵＩテスタ，第8回ヒューマンインタフェースシンポジウム論文集, pp. 287-290, 1993.

[9] 旭敏之，井関治：使いやすさ評価システム"ＵＩテスタ"の提案，情報処理学会ヒューマンインタフェース研究会資料, HI-38-1 (1991).

[10] Asahi, T., Okada, H., Iseki, O.: Tools for Iterative User Interface Design: UI-tester and OST, Proc. HCI International '95, pp. 381-386, 1995.

[11] 岡田英彦，大津祐司，旭敏之，井関治：ＵＩテスタにおける共通誤対話の分析，情報処理学会ＨＩ研究会報告, HI-50-4, pp. 25 - 31, 1992.

[12] 岡田英彦，旭敏之，井関治：ＵＩテスタによる共通操作時間モデルの分析，第10回ヒューマンインタフェースシンポジウム論文集, pp.649-652,1994.

[13] 旭敏之，岡田英彦，井関治，松田良一：ユーザの操作履歴をもとに，使いにくさを出荷前に改善，日経エレクトロニクス, No. 609, pp.111-120, 1994.

[14] Nielsen, J., Molich, R.: Heuristic Evaluation of User Interfaces, Human Factors in Computing Systems, CHI '90 Proceedings, pp. 249-255, 1990.

[15] Nielsen, J.: Finding Usability Problems Through Heuristic Evaluation, . Human Factors in Computing Systems, CHI '92 Proceedings, pp. 373-380, 1992.

[16] Nielsen, J.: Enhancing the Explanatory Power of Usability Heuristics, CHI '94 Proceedings, pp. 152-158, 1994.