

古写経翻刻のための採用支援機能の構築

福岡 整¹ 村川 猛彦² 中川 優²

¹和歌山大学 大学院システム工学研究科 ²和歌山大学 システム工学部

バージョン管理システム Subversion でデータを管理し、金剛寺一切経の撮影画像と大正新脩大蔵経のテキストファイルを対照表示させ、経典画像と同一内容のテキストを効率よく作成すること（翻刻）を支援するシステムを開発している。本稿では、仏典の専門知識を持つ者が行単位で編集できる機構を構築したので報告する。アコーディオンメニューを用いたインタフェースにより、複数人の作業結果を一覧で表示し、そのいずれかを選ぶことも、元のテキストを選ぶことも、編集することも可能としている。評価実験により、作業者数が多いほど修正すべき箇所の見逃しが減少することを確認した。

Decision-Aid Module for Transcription of Ancient Buddhist Sutras

Hitoshi Fukuoka¹ Takehiko Murakawa² Masaru Nakagawa²

¹Graduate School of Systems Engineering, Wakayama University

²Faculty of Systems Engineering, Wakayama University

We have been developing the system, using the version management system Subversion, for helping one to make the text files which are the same in content as the shot images of Buddhist sutras of Kongoji Temple while the early documents derive from the text dataset of Taisho Tripitaka. In this paper, we report the feature of decision aiding which we implemented so that an expert on Buddhist study can edit the lines smoothly. Thanks to the interface with the accordion menu, the contents redacted by operators are listed, and the decider can choose one, pick the original line, or freely change the line. The evaluation experiment showed that the oversights decreased with an increase in the operators.

1. はじめに

天野山金剛寺所蔵の一切経（金剛寺一切経）は、国内古写経の内容理解において重要な経典となっている。これまで、国際仏教学大学院大学の研究者らにより、高精細デジタルカメラを用いて悉皆撮影が行われてきた[1]。筆者らは、その撮影画像とテキスト情報を対応付けるための手法の確立を目指している。

視認性に優れる膨大な経典画像を検索するためには、経典画像と同一内容のテキストが必要であるが、既存の網羅的な漢訳仏典テキストは、大正新脩大蔵経をもとに作成されているため、古写経と異同が見られる。そこで、バージョン管理システム Subversion を用いて、複数人による作業で異同を効率良く発見、管理できるシステムを開発してきた[2][3]が、正確なテキスト作成を行うには、専門知識の持たない作業者だけでは不十分であり、専門家が妥当性をチェックする必要がある。

本稿では、この問題を解決するために構築した採用支援機能について報告する。

2. 準備

2.1 漢訳仏典とその分類

本研究が対象とする漢訳仏典は、金剛寺一切

経と大正新脩大蔵経の 2 種類である。これらは流通経路、電子化方法がそれぞれ異なっている。金剛寺一切経は、漢訳された仏典が中国への修行僧の書写を通じて持ち帰られ、国内で流布していったとされる。仏教学、国文学などにおいて特に重要とされており、近年、その存否状況の調査と、デジタルカメラを用いた経典撮影（電子画像化）が座主の協力のもと行われてきた。金剛寺一切経の経典撮影画像は 3 千巻を超え、その中のいくつかの経典はテキスト化が試みられているが、網羅的なテキスト作成には至っていない。

仏教学において漢訳仏典を引用する際に広く用いられる大正新脩大蔵経は、高麗版を中心とした刊本一切経を底本としている。刊本一切経の刊行時期は上述の古写経より少し遅れて 10 世紀以降とされている。大正新脩大蔵経については、日本の大蔵経テキストデータベース研究会 (SAT)[4][5]、および中華電子仏典協会 (CBETA)[6]により、そのテキスト化が行われ、約 9 千巻にも及ぶそのテキストデータは無料で検索および入手が可能である。

2.2 翻刻支援機能

金剛寺一切経の経典画像から対応するテキストファイルの特定を支援するため、筆者らはこ

れまで、全文検索システム[7]を構築した。しかし、金剛寺一切経のテキストデータが存在せず、大正新脩大藏経のテキストデータで代用しているため、2つの經典間の異同（文字単位の違い）により、検索してもヒットしないといった問題がある。そこで、筆者は新たに仏典テキスト校訂支援システム[3][8]を開発した。なお、現在では採用支援システムではなく翻刻支援機能と名称を変更している。そもそも校訂とは、『あくまで原作者の作品に対して忠実であろうとする立場にたち、その原文の復元を最終目標とする』[9]という意味であるが、本研究では、原文の復元を目指すことは想定していない。どちらかと言えば、翻刻すなわち『古文書、写本など崩し字で書かれた文献を楷書に直して一般的に読める形式にすること』[10]の意味が適切であると考え、名称変更に踏み切った。

この翻刻支援機能では、複数の作業者が、經典画像とテキストを照合しながら、作業者に異同を見つけてもらい、經典画像に合致したテキストを作成できる。画面例を図1に示す。



図1: 翻刻支援機能の画面例

左の画像は金剛寺一切経撮影画像であり、右のテキストは大正蔵のテキストデータである。經典画像をマウスでドラッグするかページ送りのボタンをクリックすると、右の經典テキストデータが經典画像と連動して移動する。

作業者がもし異同を見つけた場合、ボタンを押せばテキストボックスが出現するので、そこで適切な文字に修正する。そして、書き換え完了のボタンを押すとサーバに送られ、画面上では、どこを修正したか分かるよう、右の經典テキストの修正した行に色付けがなされる。

この機能のもう一つの特徴として、複数の作業者が同一文書を見て別々に修正できる点が挙げられる。これは、分量の多い漢訳仏典に対して、一人の作業者によってミスや漏れがなく校

訂するのは困難であることを想定したためである。各作業者の作業前後のテキストデータの管理はバージョン管理ソフトウェアのSubversion[11]を利用するが、今回は複数人での作業であるため、衝突が起こらないように配慮する必要がある。そこでSubversionのリポジトリを作業者の人数分だけ分岐させ、各作業者は分岐されたりリポジトリでアップデート・コミットするようにしている。

2.3 採用支援機能の必要性

過去の報告[3]では、複数の作業者の作業後テキストデータをもとに、あらかじめ閾値を設けた上で、それ以上の修正が行われている箇所を採用する方針をとっていた。しかし、作業ごとに、閲覧行数や修正箇所数にばらつきが見られ、判断に迷ったときは、修正を行わないもいた。また、經典に記されている意味や語句などを知らない作業者にとって、經典画像の異同を判別することはできても、実際にどの漢字が適当であるかを断定するのは困難である。以上より、閾値による一律で機械的な採否方法では十分な翻刻が行えないと判断し、専門家が複数人の作業者に対して、修正が適切かどうかチェックを行う機構を取り入れることにした。

3. 採用支援機能

3.1 採用支援機能の概要

本稿で述べる「採用支援機能」とは、複数の作業者が翻刻支援インタフェースを利用して生成した、作業後テキストに対して、専門家が閲覧し、適切な修正を取捨選択するものである。図2は、システム利用の流れおよび採用支援機能の位置づけをまとめたものである。

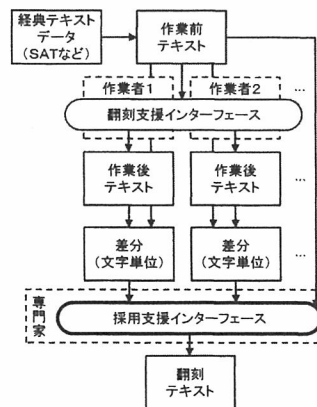


図2: システム利用の流れ

翻刻支援機能では、經典画像とテキストを照合しているが、前処理として、經典画像とテキストデータを対照表示するために、手作業で改行位置を合わせている。この状態のテキストを「修正前テキスト」と呼ぶ。各作業者は、まず翻刻支援機能にログインし、翻刻対象の經典を選択する。作業者にとって翻刻をしていない經典を選択した場合、リポジトリに分歧が作られ、作業コピーとして割り当てられる。分歧によってリビジョン番号は1増えるが、分歧元の修正前テキストの内容（およびその最後の修正のリビジョン番号）に変化はない。これにより、各作業者は共通の作業前テキストを使用して翻刻作業を行うことができる。

翻刻を終えると、作業者ごとに修正後テキストが生成されるので、修正前テキストと各作業者の作業後テキストの間で文字単位の差分を求める。異同箇所には色付けを施しており（次節）、これは、各作業者がどこを修正したかを採用支援機能で見やすくするためである。

全ての修正後テキストに対して、差分と色付け処理が終わると、採用支援機能にも色付けが反映されるので、専門家が修正前テキストと各作業者の作業後テキスト閲覧しながら、適切な修正を取捨選択していく。なお、本機能では1人の専門家が利用することを想定している。

3.2 差分データの処理

校訂支援機能では、各作業者の作業後テキストデータを並べて画面に出力するが、その際に、作業前テキストデータとの変更点に色付けを行い強調している。この強調を実現するには、各作業者の修正箇所を取得する必要があるが、Subversion における svn diff コマンドや、Linux などの diff コマンドによる差分では、行単位でしか差分を知ることができず、その行のどこを修正したのかという文字単位の取得には対応していない。そこで、自前で文字単位での差分を算出している。

具体的には、diff コマンドが行単位で取得することを利用しており、作業前・作業後テキストの文字間に改行を挿入し、保存する。次に、diff を実行し出力結果を取得し、出力結果をもとに HTML で適切に表示されるよう色付けを行い、初めに挿入した改行を削除している。

3.3 画面例

ユーザ名とパスワードで認証を行い、經典を指定すると、図3のように表示される。

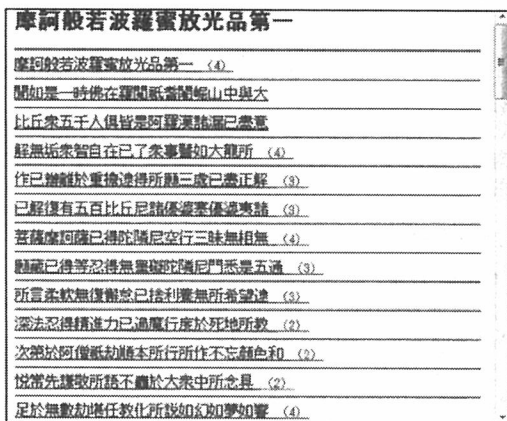


図3:採用支援機能 (初期画面)

各行は編集前テキストであり、行末の数字は、その行での修正した作業者の人数を表している。任意の行をクリックすると、その行に関する①各作業者の作業後テキスト、②編集可能なテキストボックス、③備考記入用のテキストボックスがアコーディオンメニューとして動的に引き出され、表示される (図4)。

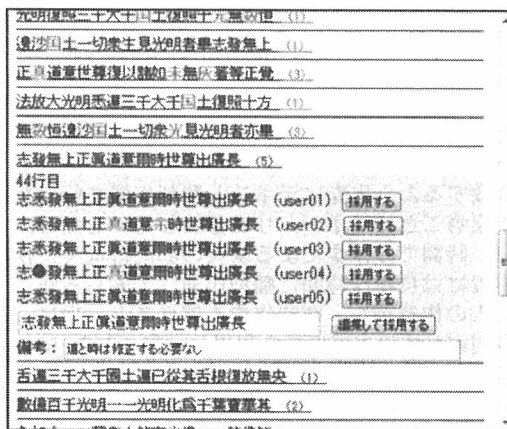


図4:採用支援機能 (行の編集)

作業後テキストは、各作業者がその行のどの文字を修正したか分かりやすいように、文字単位で、挿入・削除・置換に応じて異なる色分けを行っている。誰かの作業者の内容を採用したければ、その者の「採用する」のボタンを押せばよい。ただし、システムでは修正内容のみを管理し、誰の修正を採用したかについては管理しない。編集可能なテキストボックスは、各作業者の修正がどれも適切でない場合に、専門家が書き換えを行えるように設けている。また、備考記入用のテキストボックスには、採用した理由を入力することができる。

採用ボタンをクリックすると、その修正がサーバに送られ、コミットが行われる。画面上では、採用した内容に変更されるとともに、変更箇所には色付けがなされ（編集可能なテキストボックスで修正した場合を除く）、行末の数字が灰色になる。これにより、専門家がどこまで採用を行ったかを知ることができる。

4. 評価実験

4.1 実験方法

本機能を利用し、適切に翻刻が行えるかを確かめるために、評価実験を行った。

対象の経典は、『摩訶般若波羅蜜放光品第一』で、446行、約7,900字の分量がある。テキストはSATのテキストデータを利用した。あらかじめ、検索結果[12]から不要な文字等を取り除いた上で、経典画像と行が合うように手作業で改行位置の変更を行った。これまでシステム構築や評価実験で使用してきたCBETAのテキストと比較して、字体の違いは相当数あり、SATのテキストの方が、現代日本語に近い字形を使用している（「衆」「俱」など。）。CBETAで外字扱いされていた箇所が多くがSATでは1文字に割り当てられていたのに対し、その逆のケースは見られなかった。

実験協力者は、経典に関する知識のない大学生5人と古写経を専門とする研究者1人である。

実験手順は以下のとおりである。まず5人の大学生（以下「作業者」という）が、別々のユーザとして翻刻支援機能にログインし、画像に合致するよう実際にテキスト翻刻作業を行った。作業者ごとにリポジトリを分岐しているため、同一時間での作業による衝突の心配はなく、作業者には自由な時間に翻刻作業してもらった。5人の作業者が、翻刻作業を終えると、次に、専門家が採用支援機能を利用し、得られた5つの翻刻後テキストを参照しながら適切な修正を見つけ採用してもらった。

4.2 実験結果

この実験において、専門家は冒頭からの70行を閲覧し、9行の採用を実際に行っていた。各作業者の結果は表1のとおりである。

表1:各作業者の翻刻結果

作業 者名	修正 行数	採用 行数	(%)	編集採 用行数	(%)	見逃 し数	(%)
user01	15	2	13.3	4	26.7	5	55.6
user02	50	1	2.0	5	10.0	4	44.4
user03	4	2	50.0	2	50.0	7	77.8
user04	30	2	6.7	4	13.3	5	55.6
user05	23	2	8.7	3	13.0	6	66.7

表の各列の意味について述べる。「修正行数」は、70行までで実際に修正した行数である。これを見ると、user02が50行の修正を行っているのに対し、user03が4行ほどしか修正をしていないなど、作業者によって、ばらつきがみられる。「採用行数」は、作業者が修正し、かつ専門家がそれと同じものを採用した行数を言う。なお、3.3節で述べたとおり、誰の修正を採用したかについては管理していないので、行ごとに一致しているか否かを比較し、計数した。「編集採用行数」は、作業者が修正し、かつ専門家も修正した行数を言う。その内容が一致しているかは問わない。値は必ず、採用行数以上となる。この尺度を導入したのは、「ここは修正すべきではないか」と作業者が判断したものが、その修正内容が適切でなかったとしても、専門家の採用決定に影響を与えたことを表すためである。最後に「見逃し数」は、専門家が修正したが作業者は修正をしなかった行数であり、専門家の採用行数（本実験では9）から編集採用行数を引いたものとして計算される。

採用行数と編集採用行数のパーセンテージは、修正行数を分母としている。また見逃し数のパーセンテージは、専門家の採用行数を分母としている。個別で見ると、最も見逃し率の低い作業者でも40%台であり、多くの作業者が、半分以上を見逃していることが分かる。

そこで、5人の作業者の何人が協力すれば見逃し率がどのようになるかを算出した（表2）。

表2:想定作業者数ごとの見逃し率

想定作業 者数	組み合わせ	見逃し率 (%)
1	[1][2]...[5]	60.0
2	[1,2][1,3]...[4,5]	45.5
3	[1,2,3][1,2,4]...[3,4,5]	36.6
4	[1,2,3,4]...[2,3,4,5]	28.8
5	[1,2,3,4,5]	0.0

想定作業者数を1~5人としたとき、それぞれで、作業者の組み合わせを作ることができる。括弧内の数字は作業者名の略記であり、例えば「[1,2]」はuser01とuser02の編集結果に限定したとき、専門家の採用行数と比較して、どれだけ見逃しがあるかを求めている。表の見逃し率の値は、同一想定作業者における全組み合わせの見逃し率の平均により計算している。

今回の実験では、作業者が5人のとき、見逃し率が0%となった。また、想定作業者数が多いほど、見逃し率が減っていることを確認した。作業者数の採用に対する影響を見るため、別

の算出を試みた。表 3 は、ちょうど i 人 ($0 \leq i \leq 5$) が修正した行数とそれぞれの専門家の採用行数の関係を表している。

表 3:修正人数ごとの行数と採用率

作業人数	修正行数	専門家の採用行数	(%)
0	17	0	0.0
1	16	2	12.5
2	17	1	5.9
3	11	2	18.2
4	6	3	50.0
5	3	1	33.3
計	70	9	12.9

ここでも、修正した作業者が多い行ほど、採用される確率が高いことを確認した。なお採用支援インタフェースとしては、修正人数が 0 人の行も編集できるようにしているが、それによる修正は見られなかった。

最後に、今回の実験で実際に採用された例を示す(表 4)。

表 4:各作業者の修正例と採用例

SATのテキスト	優婆塞優	踊没諸有地	床於此三千
經典画像 			
user01	優婆塞優優	—	狀於此三千
user02	—	—	狀於此三千
user03	優婆塞優優	—	—
user04	優婆塞優優	踊没諸有地	狀於此三千
user05	優婆塞優優	—	狀於此三千
専門家	優婆塞優優	踊没諸有地	牀於此三千

例えば、SAT のテキストが、「優婆塞優」であるのに対し、5 人中 3 人が、「婆」と「塞」の間に「夷」という字を挿入し、専門家が、3 人の修正が適切であると考え、採用を行った(前後関係から、この文字の挿入は書写ミスに基づくものと推測される)。また「踊没諸有地」は、作業員 1 人が「没」を「肢」に置換しているが、専門家は「没」を「波」と判断し、テキストボックスで編集を行っている。同様に、「床於此三千」を作業員 4 人が「床」を「狀」に置き換えていたが、専門家が、「床」をその異字である「牀」に変更している。

4.3 考察

実験の結果、作業者が多くなるほど、見逃しが少なくなることを確認した。採用事例においては、多くの作業者が行った修正をそのまま採用したものや、多くの作業者が行った修正と異なる文字に置き換えて採用としたものが見られた。これらは、作業員の修正内容が、専門家の編集作業に役立っていることを意味し、複数人の作業員と 1 人の専門家が連携して、本システムで經典画像に合ったテキストを作ることができたと言える。

評価実験により得られた結果に基づき、考察を行った。翻刻作業を行った 5 人の修正をみると、ばらつきが見られるが、これには、4 つの原因が考えられる。1 番目の原因として、単なる見落としである。システムにより負担が軽減されたとはいえ、文字同士の照合は、入念に作業をしなければ、修正漏れの原因に繋がる。2 番目の原因として、作業員ごとの文字認識の違いである。作業員により、經典画像の文字とテキストの文字を照合した時に、同じであると考える者もいれば、違うと考え修正を行う者もいる。3 番目の原因として、作業意識の違いである。經典には、多くの異体字など読みづらい文字が存在し、日常的にコンピュータを使う者にとって、修正すべきか判断できないという事態が起り得る。そこで、無理をして修正を行う者もいれば、修正しない者もいる。4 番目の原因として、修正しなくてよい文字を複数回に渡り、修正してしまうケースである。実験においては、user02 と user04 が、専門家が修正する必要がないと判断した「所」という文字を 16 箇所修正している。この状況では、作業員の作業効率低下と専門家の負担に繋がる可能性があり、今後、検討する必要がある。

ここで、テキスト修正作業の直列・並列について考察を行う。

直列型というのは、誰かが修正した後で、他の誰かが修正していくという形態をいう。SAT のテキストデータベースは、[13]を読んだ限り、直列型である。また並列型というのは、修正を複数人で同時に行う形態をいう。本稿および[3]は並列型と言える。なお、Subversion の使用や Web アプリケーションとしての構築は、並列型であることを直には意味しない。例えば、[2]のシステムは、衝突を起さないよう時分割して同一經典テキストの編集をさせていたので、直列型と言える。

直列型の特徴を述べる。まずメリットとして、実装が単純であることが挙げられる。しかしデメリットも数多く指摘でき、もっとも深刻な問

題は、複数人の作業者をめぐって、特定箇所の頻繁な修正 (A→B→A→B→…) が起こり得る点であろう。これについては、そのような修正を検知する仕組み (結局のところ、履歴管理) を設けておくか、そのようなループにならないよう、作業者間で合意を得ておく必要がある。他に、どのような状態になれば (または、どれだけ時間をかければ) 終了とすればよいかの判定が困難という点も、課題となる。これは、どれだけバグを取り除けばソフトウェアを納品できる水準になるかという、ソフトウェア工学の観点に通じるところがある。

次に、並列型の特徴を記す。直列型と比べて、分岐または何らかの衝突回避手段をシステムで持つ必要があるという点で、実装が複雑になる。本研究においては、リポジトリ URL の命名規則と作業コピーのディレクトリ階層を適切に定め、分岐を `svn branch` コマンドで実施し、通常のバージョン管理を含むソフトウェア開発では不可欠な、`svn merge` を使用しないことで、小さな手間で並列編集を可能とした。作業者ごとに別々に翻刻作業を行うので、修正のループは考慮しなくてよい。履歴管理に関しては、1行の修正により 1 回コミットというポリシーにより、いつ誰がどこをどのように修正したかを容易に知ることができる。加えて、(不) 採用数や (不) 採用率などを定量的に表せるため、どの作業者が有能かを数値化すること、典型的な編集間違い例を抽出・提示することなどが期待できる。終了までの作業時間については、翻刻支援機能を使った編集ではそれまでの実績をもとに所要時間を決めて実施し、採用支援機能の使用においては、一人でも修正した行の点検が終わるまでアクセスすればよい。

5. おわりに

本稿では、翻刻支援機能、採用支援機能およびその評価実験について述べた。本システムにより、複数人の作業者と 1 人の専門家が連携することで、修正漏れの少ないテキスト翻刻ができることが期待できる。

今後の課題としては、採用支援機能の画面内で、行に対応する経典画像が閲覧できるインタフェースの開発などが挙げられる。

謝辞 実際に採用支援機能をご使用いただき教え切れないアドバイスを賜りました。国際仏教学大学院大学附置日本古写経研究所非常勤研究員 三宅徹誠氏に感謝申し上げます。翻刻支援機能を使用して編集を行った学生諸氏に感謝します。

参考文献

- [1] 落合俊典: 金剛寺一切経の総合的研究と金剛寺聖教の基礎的研究 平成 16~18 年度科学研究費補助金基盤研究(A)研究成果報告書, 課題番号 15202002, 第 1 分冊 (2007).
- [2] 丁敏, 村川猛彦, 福岡整, 中川優: Subversion を用いた仏典テキスト校訂支援システム, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, Vol.2008, No.15, pp.61-66 (2008).
- [3] 福岡整, 村川猛彦, 野田大地, 中川優: Subversion を用いた仏典テキスト校訂支援システムの評価, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, Vol.2009, No.16, pp.61-66 (2009).
- [4] SAT, 大正新脩大藏経テキストデータベース, <http://21dzk.l.u-tokyo.ac.jp/SAT/>
- [5] 下田正弘, 永崎研宣: 大藏経と人文系データベース, 情報処理学会研究報告 人文科学とコンピュータ, 2009-CH-82, pp.1-6 (2009).
- [6] CBETA 中華電子佛典協會, <http://www.cbeta.org/>
- [7] 村川猛彦, 丁敏, 中川優: 仏典全文検索システムの構築と評価, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, Vol.2007, No.15, pp.221-228 (2007).
- [8] Murakawa, T., Fukuoka, H., Noda, D. and Nakagawa, M.: Transcription Support System using Subversion, 6th International Conference on Web Information Systems and Technologies, Valencia, Spain, Vol.2, pp.150-155 (2010).
- [9] 秋山陽一郎: 校訂とはいかなる行為か?, 漢字文献情報処理研究, Vol.6, pp.36-43 (2005).
- [10] 翻刻とは - はてなキーワード, <http://d.hatena.ne.jp/keyword/%CB%DD%B9%EF>
- [11] Subversion, <http://subversion.tigris.org/>
- [12] http://21dzk.l.u-tokyo.ac.jp/SAT/T0221_08,0001a01:0221_08,0007b07.html
- [13] 永崎研宣, 鈴木隆泰, 下田正弘: 大正新脩大藏経テキストデータベース構築のためのコラボレーションシステムの開発, 情報処理学会研究報告 人文科学とコンピュータ, 2006-CH-70, pp.33-40 (2006).