

機械翻訳を介したオープンソースソフトウェア開発： Intercultural Collaboration Experiment 2002

野村 早恵子[†] 船越 要[‡] 山下 直美[‡] 安岡 美佳[†] 石田 亨^{†*}
 京都大学大学院情報学研究科社会情報学専攻[†]
 日本電信電話(株) NTT コミュニケーション科学基礎研究所[‡]
 科学技術振興事業団^{*}

1 はじめに

アジア太平洋地域のインターネット人口の割合が急激に伸び、2003年にはアメリカ(1億6280万人)や欧州(1億6220万人)を追い抜き世界トップの1億8330万人になると予測される^[1]。北米や欧州で急速に活発化している地域内連携に呼応するかたちで、アジア地域での国際共同プロジェクトを増やしていく必要がある。

しかしアジアのように、言語が全く違う国により構成される地域にとっては、言語の壁が共同プロジェクトの大きな問題となりうる。アジア各国の言語教育の現場では、隣接している国同士の言語についてほとんど教育されず、互いに相手の言葉を理解することが難しい。また、アジアの有名大学の学生に尋ねると、チャットや論文は英語で書くが、プロジェクトの資料を英語で書くことには大きな抵抗があることが分かる。思考が母国語で行われているため、プロジェクトの中間資料は母国語で書きたいのである。このような状況認識に立ち、本研究ではアジアでの共同プロジェクトを支援するために、言語の違いを超えるための異文化コラボレーション支援システムを構築する。

インターネットを介したコミュニケーションでは、対面でのコミュニケーションとは違い機械翻訳技術を簡単に利用することができるため、技術の向上が異文化コラボレーションを飛躍的に伸ばす可能性を持つと考えられる。これまでの機械翻訳は、一文単位の翻訳が中心であり、新聞記事やマニュアルなど書き言葉を対象として研究が行われている^(2)[3][4][5][6])。研究開発の評価は一文単位の翻訳の正確さにある。一方、我々が本研究において重要と考えることは、一文一文が正しく訳されることではなく、相手にいかに正しく意図を伝えることができるかである。そこで、本研究では実証実験を通じて、「共通の目標をもつ構成員のコラボレーション過程」に機械翻訳を用いる文脈を特定し、異文化間対話における機械翻訳の可能性を模索する。

本稿では、機械翻訳を介したコミュニケーションでの言語現象を細かく分析する^[6]のではなく、ノイズの高いメディア上で、いかにして異文化コラボレーションが達成されたかを考察する。

2 異文化コラボレーション実験 (Intercultural Collaboration Experiment 2002: ICE2002)

言語の壁を超えた異文化コラボレーションの可能性を模索すべく、本研究では、異文化コラボレーション実験(Intercultural Collaboration Experiment 2002 (ICE 2002))を行った。実験には、京都大学、上海交通大学(中国)、ソウル国立大学およびハンドン大学(韓国)、そしてマラヤ大学(マレ

シア)の計算機科学に関連した学部および研究科の教員とその学生総勢約40名が参加した(図1参照)。

異文化コラボレーション実験では、オープンソースソフトウェア開発が行われたが、その際、実験参加者は、一度も対面することなく、ネット上でバーチャルチーム^[7]を構成し、ソフトウェア開発を行った。



図 1: ICE2002 の参加者

2.1 実験スケジュールと参加者の任務

ICE2002は、第1トラック(5月~7月)、そして第2トラック(9月~11月)と、それぞれ約2ヶ月間の日程で行われた。参加者の情報交換を活発に行わせるための工夫として、各トラックは各4週間ずつのソフトウェア設計フェーズとソフトウェア開発フェーズに分けるかたちで構成された。

前半4週間の「ソフトウェア設計フェーズ」(第1フェーズ)は、各参加者が開発したいソフトウェアに関する意見をそれぞれ出し合い、それに対して互いにコメントをしよう期間である。後半4週間には、ソフトウェア設計フェーズで提出された設計案の実装を行う「ソフトウェア開発フェーズ」(第2フェーズ)が設けられた。このフェーズでは、ソフトウェア開発のためのバージョン管理システム(CVS)が運営サーバに開設され、分散環境におけるソフトウェア開発の効率化がはかられ、それぞれのモジュールを最終的にひとつのソフトウェアに統合するという目標が立てられた。

2.2 多言語コラボレーションツール

実験では、参加者間の日常のディスカッションに用いるツールとして、TransBBSおよびTransWEB(図2参照)が提供された^[8]。両ツールは、arcnet/sangenjaya¹(NTT MSC Sdn.Bhd.)およびJ-server²((株)高電社)のインターネット

Open Source Software Development Using Machine Translation: Intercultural Collaboration Experiment 2002

Saeko Nomura[†], Kaname Funakoshi[‡], Naomi Yamashita[‡], Mika Yasuoka[‡], and Toru Ishida^{†*}

[†] Department of Social Informatics, Kyoto University

[‡] NTT Communication Science Laboratories,

^{*} Japan Science and Technology Corporation

¹ <http://sangenjaya.arc.net.my/index-e.html>

² <http://www.j-server.com/index.shtml>

経由での翻訳サービスを介し、中国語、日本語、韓国語、マレー語、そして英語間のテキスト翻訳を行うコミュニケーションツールである。

TransBBS は、開発ソフトウェアの設計に関する意見交換や開発状況の報告など、実験参加者同士の日常的なやり取りに用いられた。一方、TransWEB は、HTML 形式のソフトウェア開発ドキュメントを母国語に翻訳して閲覧するために提供された。特に効率的な分散型ソフトウェア開発には、知識共有のためにシステム構成図やデータフロー図などが入った文書が必要である。各国の参加者は、母国語で作成されたソフトウェア開発ドキュメントを HTML 形式で保存し、TransBBS 上でドキュメントが提出された URL を他の参加者に伝える。それを受けて、他国の参加者は、TransWEB 上でドキュメントを母国語に翻訳して参照する。このように TransBBS と TransWEB は、ソフトウェア開発ドキュメントを介した、グループ間の情報共有のために利用された。

なお、実験期間中は、国内、国外を問わず、ディスカッションはすべて TransBBS および TransWEB 上で行うことがルール付けられ、電子メールや Chat システムなどを用いたコミュニケーションは禁止された。

3 多言語環境におけるコミュニケーションの観察結果

実験では、各国の参加者は、以下のツールを開発した。日本はエージェントが多国籍チームに参加するメンバーのスケジュール管理を行う“TransGroupware”，中国は多言語情報検索を可能にする“TransSearch”，韓国は多言語同期会話ツールである“TransChat”，そしてマレーシアは主にモバイルユーザのための多言語ショートメッセージサービスである“TransSMS”を中心に開発した。これらのツールの特徴は、機械翻訳を介した異文化コミュニケーション実現のための機能を有しているところにある。最終的には、これらはユーザデータベースの統一など、ひとつのインタフェースに統合された形で提出された。

実験過程では、機械翻訳技術を介した参加者間のコミュニケーションのあり方が、ソフトウェア設計フェーズとソフトウェア開発フェーズで以下のように異なっていることが観察された。

- ソフトウェア設計フェーズ: 社会的メッセージ⁹⁾のやり取り。利用システムに関する質疑、翻訳の質に関する意見の投稿や翻訳の質を改善するためのやり取り。
- ソフトウェア開発フェーズ: 議論の内容が、ソフトウェア開発に直接関連するトピックのみへと収束。翻訳の質を問わず積極的に行われた技術的な質問とそれに対する反応。

そこで本章では、これら 2 つのフェーズで観察された参加者の行動を通じ、機械翻訳機能が異文化コラボレーションの文脈において利用可能かどうかを考察する。

ソフトウェア設計フェーズでは、各参加者による自己紹介からはじまり、翻訳の質を確認するためのテスト投稿、翻訳の質に関する意見や互いの投稿に対する意味の確認、また TransBBS の機能などに関する質問などが投稿された。

前半においてよく観察された翻訳の質に関する議論では、「他のグループの参加者が投稿したメッセージの意味がわからない」などという不満メッセージが投稿された。そのような中、実験に積極的に参加していたものは、自分



図 2: ICE2002 で用いられたコミュニケーションツール: TransWEB(左)および TransBBS(右)

の投稿メッセージに対する翻訳の誤りについて積極的に対応し、他のグループとのコラボレーションをはかろうとする行為が観察された。例えば、次のメッセージは、日本からの参加者が、翻訳サービスに注意を払わずに投稿をすると、他の国の参加者が理解しづらくなることを示している。

「日本チームの Discussion ルームでは、どの程度翻訳精度を気にかけるべきですか？日本語で通じてしまうと思うと、思わずそのまま投稿してもよいのではと感じてしまうのですが、他のチームから何を議論しているのかわからないのは問題ですね。気をつけるようにします。」

このようなノイズの高いメディアを用いる中で、参加者が相手に正しく意図を伝えるために行った翻訳誤りリペア行為は、二つの種類に分類することができる。

- 自己主導型リペア(Self-initiated repair): TransBBS に投稿する前に、翻訳の質をあげるために何度も翻訳精練行為を自主的に行う行為
- 他者主導型リペア(Other-initiated repair): メッセージの読み手が意味の確認を行ったり、意味がわからないというリアクションを行ったりすることにより、翻訳の精練作業がインタラクティブに行われるもの

TransBBS 上に投稿されたメッセージが、投稿前に何度自主的に翻訳の精練作業を行ったのか、つまり自己主導型リペアの頻度を、表 1 に表す。

表 1: TransBBS における自己主導型翻訳リペア

ソフトウェア設計フェーズ		ソフトウェア開発フェーズ	
Name	# post / translate before posting	Name	# post / translate before posting
Saeko	69 / 234	Tomko	60 / 294
Bikesh	66 / 112	Saeko	43 / 101
Tomoko	51 / 395	LimSanny	24 / 68
Yoshiyuki	35 / 170	HaoLei	19 / 22
LimSanny	28 / 71	KangWon	18 / 56

最も自己主導型リペアを行った参加者は、実験開始直後であるソフトウェア設計フェーズには、平均すると、1 つの投稿メッセージにつき平均約 8 回のリペアを行っている。翻訳精練作業の一例を以下に示す。

初期投稿	私の文章もあまり上手に翻訳されません。日本語の勉強から始めようかな。	My sentence is also and it is not translated well. Will it begin from the study of Japanese?
5回の翻訳リペアの後...		
最終投稿	私の文章も上手に翻訳されません。私はもう一度日本語の勉強を始める必要があります。	My sentence is not translated effectively either. I need to begin the study of Japanese once more.

このように、実験参加者は、翻訳ツールを利用する過程で、うまく翻訳される文章の構造を模索し、できるだけ単純かつ文語体のメッセージを書くように心がけるようになった。また、複数言語を操ることができる中国からの参加者が、機械翻訳を利用するにつれ正しく翻訳されるためのノウハウを習得し、それを他の参加者に報告する動きも見られた。

中国語(再翻訳)	日本語	英語(再翻訳)
能得到如果以下的事注意制作文章源于利用自动翻译公告牌投稿的时候, 翻译结果。我那样思念。1. 主语和宾语, 不省略谓语。2. 句读点也确实放入。3. 文章尽可能逐地输入。4. 俗话和惯用语, 省略语不使用。5. 汉字事多, カタカナ言词カタカナ输入。平假名事多的话, 有很好地不被翻译的情况。(日語的场合)6. 有人物的名字等的专有名词不能翻译的事。尽可能, 专有名词英语记教也附记。Tomoko, 怎样想?	自動翻訳を利用して掲示板に投稿する際に、以下のことに注意して文章を作成すればよい。翻訳結果が得られません。私はそう思います。1. 主語と目的語、述語を省略しません。2. 句読点も正しく入れます。3. 文章は出来るだけ短く入力します。4. 俗語や慣用語、省略語は使いません。5. 漢字を多用し、カタカナ言葉(カタカナで入力)します。ひらがなを多用すると、うまく翻訳されない場合があります。(日本語の場合)6. 人物の名前などの固有名称は翻訳できず、事があります。できるだけ、固有名称には英語表記も付記します。Tomoko, どう思いますか?	In contributing using automatic translation to a notice board, if paying attention to the following things, it makes a sentence, a better translation result is gained. I think so. 1 A subject and an object and a predicate are not omitted. 2 Punctuation marks can also enter accurately. 3 Sentences are input as short as possible. 4 A slang and an idiom and a clipped word are not used. 5 Using often a Chinese character, it inputs a katakana word by katakana. If a Japanese syllabic alphabet is used often, it may not be translated effectively. ----- 6----- I also append the English notation to a proper noun as much as possible. Tomoko -----

一方、他者主導型のリペアでは、多国籍間で翻訳の意味確認が行われた。次に示す会話例では、他国に住むもの同士が翻訳の精練作業を強調して行っている。図の太線で囲まれたメッセージは、投稿された原言語を表す。この例では、日本人の参加者が挨拶の場で「よろしくお願ひします」という日本語独特の挨拶文を投稿している。この「よろしくお願ひします」は、会話のコンテキストにより意味が異なるという、翻訳しづらい文章である。ここでは、誤って中国語に訳された「よろしくお願ひします」の本来の意味を確認するためのやり取りが行われ、結果的に多言語環境で正しい知識の共有を達成している。

japan-53 Posted by: lizawa isao document Subject: Hello, World --> Date Posted: Monday, May 20, 2002 16:46:16			
Japanese	English(retranslate)	Chinese(retranslate)	Korean
初めて書き込みます。日本チームの翻譯です。よろしくお願ひします。	I write for the first time. It is the -- of its Japanese team. It is asked properly.	初次写上。是日队的翻译。适当地请求。	처음입니다. 일(奉)입니다.
china-84 Posted by: Hao Lei document Subject: Date Posted: Tuesday, May 21, 2002 12:26:47			
私は見積もって、日本の友達に言った"適切に願ひする"、"よろしくお願ひします"意味か?	I estimate. The friend of Japan said "Suitably request". Be exactly "Please concern more" Meaning?	我估计,日本朋友说的"适当地请求",就是"请多关照"的意思?	
china-86 Posted by: dingpeng document Subject: Date Posted: Tuesday, May 21, 2002 13:02:45			
恐らくそうでしょう	Probably is	大概是吧	아마 그렇지요
japanb-88 Posted by: MIKI Takeru document Subject: That's right Date Posted: Tuesday, May 21, 2002 16:52:01			
> China-84 - China-86 おしゃるとおりです。Hao Leiさんが推測した意味の日本語訳を読みました。それは、最初、に翻譯さんが書いた日本語に一致していました。	> China-84 - China-86 It is as it says. I read Japanese translation of the meaning which Hao Lei supposed. It agreed with Japanese which I wrote first.	> China-84 - China-86 是说的那样。Hao Lei读了推测的意义的日语译。那,在最初时翻译写的日语相符。	

また、多言語環境には限定されないような、つまり翻訳の質を問うものではない他者主導型のリペアも観察された。例えば、下の会話例はそのひとつである。ノイズの高いメディアを介しても、明確な質疑応答が達成されている。

中国語	日本語(再翻訳)
翻译成英语时, 出现了错误。	英語を訳す時、誤りが現れた。
中国語(再翻訳)	日本語
Donghui, 那个问题, 是在 TransSearch 上的问题? 再稍微详细地请教。如果那样做, 我们也不能解决。察知。谢谢。:-)	Donghui, その問題は、TransSearch 上での問題ですか? もう少し詳しく教えてください。そうすれば、私たちも解決できるかもしれません。ありがとう。:-)
中国語	日本語(再翻訳)
Saeko, 对不起, 我说的意思不是很明确。我是说 TransBBS 上的翻译问题。有时候, 翻译成英语的时候, 出现 Translation Error (Time out) 的问题。不过, 按“再翻译”, 翻译就成功了。可能是 Timeout 设定的问题吧。我投稿设定的时间为 30 秒。	Saeko, すみません、私の言った意味はとても明確ではない。私は TransBBS 上の翻訳の問題を言っているのだ。時には、英語を訳す時、出現 Translation Error (Timeout) の問題。でも、“更に訳す”を押して、翻訳は成功した。可能性は Timeout の設定する問題だろう。私が設定する時間投稿するのは 30 秒だ。
中国語(再翻訳)	日本語
你好。Lin, 翻译有时花费时间。这, 是翻译サーバの問題。错误, 好象做翻译的时间带也依存。我, 听说混乱的时候错误发生容易。	こんにちは。Lin, 翻訳に時間がかかることがある。これは、翻訳サーバの問題です。エラーは、翻訳をする時間帯にも依存するようです。私は、混雑しているときにエラーが発生し易いと聞きました。
中国語	日本語(再翻訳)
谢谢, 我明白了。	ありがとう、私は分かった。

このように、ソフトウェア設計フェーズでは、自己主導型リペアや他者主導型リペアが繰り返されながら翻訳の質を改善したり、参加者同士の理解促進が深まったりしたといえる。

一方、ソフトウェア開発フェーズでは、社会的メッセージのやり取りやシステムに関する投稿はほとんどなくなった。また、投稿されたメッセージに対する翻訳の質はソフトウェア設計フェーズと比べ変わらないのにもかかわらず、翻訳の精練という文脈では相手の投稿に対する意見はほぼ発せられなくなった。その一方で、ソフトウェア開発に関連するトピックでは、多国籍間での意見交換がソフトウェア設計フェーズに比べむしろ積極的に行われた。これは、特に積極的に参加しているメンバーが、テクニカルかつ具体的な内容の発言に対して、翻訳結果が完全でなくても意味を推理し、何らかのレスポンスを与えていたことに起因している。

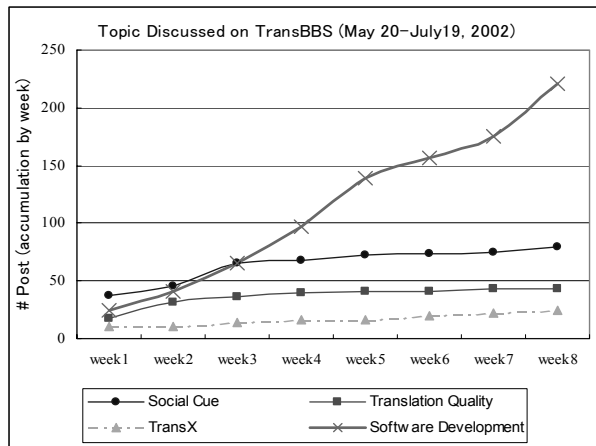
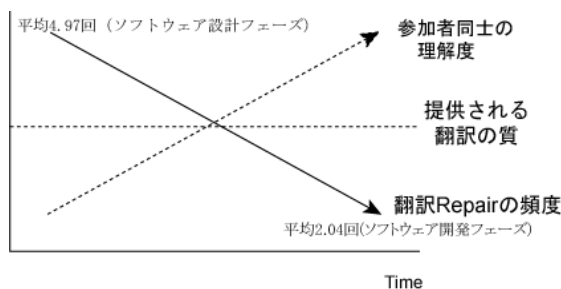
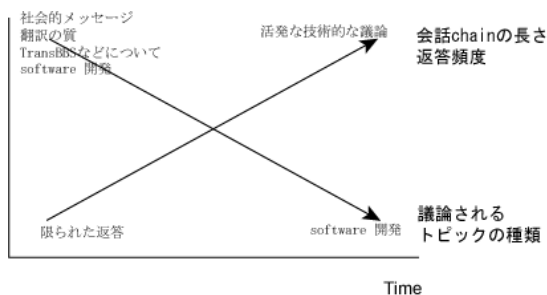


図 2: 実験過程における TransBBS 上でのトピックの推移



(a) 参加者の翻訳サービスへの適応



(b) コミュニケーションパターンの変化

図 3: 時系列に見た多言語コミュニケーションのあり方

図 2は、第 1 トラックで行われた TransBBS 上におけるメッセージ投稿をトピックごとに集計し、その推移(週ごとの投稿数の累計)を表したものである。このグラフからもわかることは、実験初期には社会的メッセージ^[8]のやり取り。利用システムに関する質疑、翻訳の質に関する意見の投稿が、ソフトウェア開発に関する議論以外に多くの割合で見られたが、実験が進むにつれ、ソフトウェア開発に特化した活発な議論がなされたことである。

以上のことをまとめると、図 3のように表すことができる。実験を通じて翻訳サービスの質は変わらなかったにもかかわらず、参加者は翻訳の能力に順応し、それを学習することにより、翻訳の精練作業の頻度を落とすことに成功した。また同時に、相互に投稿メッセージへの理解を高めた(図 2 a)。さらに、実験が進むにつれ、議論されるトピックが本来の目標追求のためのものに収束し、積極的な議論が達成された(図 2 b)。

4 おわりに

ICE2002 により明らかになったことは、以下のとおりである。共通の目標を持ったコラボレーションの文脈であれば、人は自らを翻訳の能力に適応させる。また、メッセージの読み手も、複数の言語を確かめたり、メッセージ投稿者に対して意味を確認したりしながら、他言語圏に住むコラボレーターとの理解を深めるよう努力する。特に、特定のテクニカルな知識を共有する者同士のコラボレーション過程であれば、翻訳の誤りはある程度容易に克服され、目標を達成することが可能となる。

今後、我々は翻訳機能を社会的なエンティティとしてとらえ、多国籍プロジェクトチームの中で活動できる翻訳エージェントの可能性を追求する。翻訳エージェントは、チーム内の協調型翻訳の精練をサポートし、翻訳の誤りを低減させる“翻訳精練機能”と、翻訳の質を超えたコミュニケーションを達成させる“コラボレーション機能”の二つを兼ね備える。これら二つの機能には、プロジェクトで用いられる用語やチームメンバーのアクションのオントロジーを用いるという共通点がある。

学習機能を備えた翻訳エージェントは、チームメンバーの投稿メッセージを、あらかじめ登録されたプロジェクトに関連する用語のオントロジーをもとに変換・翻訳し、またメンバーのアクションを同定する。そして、アクションの同定により、協調を促すための的確なメッセージを追加する(図 4 参照)。

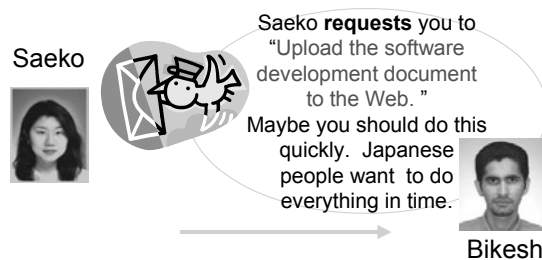


図 4: 翻訳エージェントのイメージ

今後の我々の研究課題は、他のアジアの国々との共同研究として翻訳エージェントを開発し、それを実際の仕事の場で多国籍プロジェクトを支えるツールにすることである。

謝辞 本研究は、京都大学大学院情報学研究科社会情報学専攻、日本電信電話(株) NTT コミュニケーション科学基礎研究所、科学技術振興事業団戦略的基礎研究推進事業(CREST)の主催で行われた。実験には、京都大学、上海交通大学(中国)、ソウル国立大学およびハンドン大学(韓国)、マラヤ大学(マレーシア)が参加した。実験準備および実験中においては、NTT MSC Sdn.Bhd. (arcnet/sangenjaya)の小寺博氏、阿部明典氏、(株)高電社、NTT コミュニケーション科学基礎研究所の大山芳史氏、真鍋義文氏、NTT サイバースペース研究所の林良彦氏、小倉健太郎氏、(株)数理システムの山本晃成氏、京都大学石田研究室の岡本昌之氏、中塚康介氏に多大なるご協力を得た。この場を借りて深謝いたします。

参考文献

- [1] M. Pastore, "Asian Internet Market to Surpass United States," CyberAtlas, August 7, 2001, http://cyberatlas.internet.com/big_picture/geographics/print/0,,5911861481,00.html
- [2] W. J. Hutchins and H. L. Somers, An Introduction to Machine Translation, Academic Press, 1992.
- [3] 長尾真 編, 自然言語処理, 岩波講座ソフトウェア科学 15, 岩波書店, 2001
- [4] D. Jurafsky and J. H. Martin, Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, 2000.
- [5] B. Maegaard Ed., Proceedings of MT Summit VIII, 2001.
- [6] 小倉健太郎, 林良彦, 野村早恵子, 石田亨, "目的指向の異言語間コミュニケーションにおける機械翻訳の有効性の分析-異文化コラボレーション ICE2002 実証実験から-, 情報処理学会第 65 回全国大会, 2T6-4, 2003.
- [7] S. L. Jarvenpaa, D. E. Leidner, "Communication and Trust in Global Virtual Teams," JCMC 3 (4) June 1998.
- [8] 船越要, 山本晃成, 藤代祥之, 野村早恵子, 石田亨, "異文化コラボレーション支援システムの設計," 情報処理学会第 65 回全国大会, 4A-5, 2003.
- [9] F. Henri, "Computer Conferencing and Content Analysis," A. R. Kaye (Eds.) Collaborative Learning through Computer Conferencing: The Najaden papers, pp.115-136, Springer, 1992.