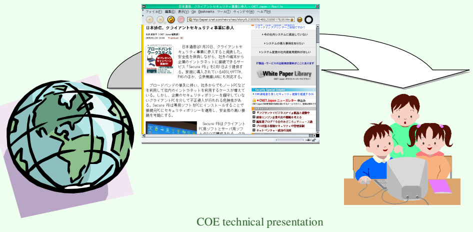
< COE Technical Presentation, 27/01/2005 >

A Category-based Approach to Paraphrase Corpus Construction

Atsushi FUJITA
Computational Linguistics Lab.
atsush-f@is.naist.jp

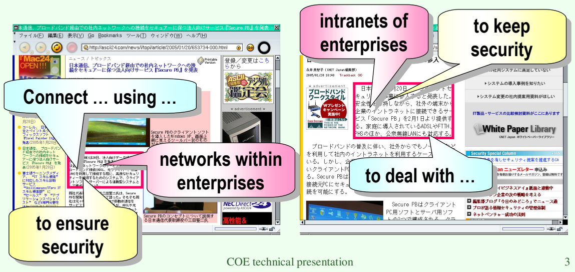
Background

- Needs of accessibility to Web documents
  - Most information is conveyed by text
- Towards intelligent treatment of text data
  - Natural Language Processing (NLP)



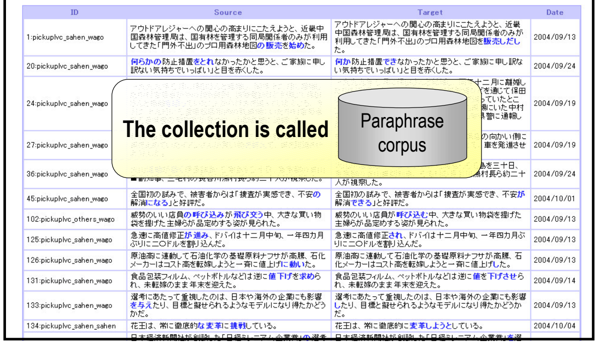
Towards intelligent content processing

- Paraphrases
  - Same information is reported using different wording
  - Handling them is useful for summarization, QA, etc.



Today's topic

- Collecting paraphrase examples



ID	Source	Target	Date
1:pickuplvc_sahen_wago	アウトドアレジャーへの関心の高まりにこたえようと、近畿中国森林管理局は、国有林を管理する同局関係者のみが利用してきた「門外不出」のプロ用森林地図の販売を始めた。	アウトドアレジャーへの関心の高まりにこたえようと、近畿中国森林管理局は、国有林を管理する同局関係者のみが利用してきた「門外不出」のプロ用森林地図を販売しだした。	2004/09/13
20:pickuplvc_sahen_wago	何らかの防止措置をとれなかったかと思うと、ご家族に申し訳ない気持ちでいっぱい」と目を赤くした。	何か防止措置できなかったかと思うと、ご家族に申し訳ない気持ちでいっぱい」と目を赤くした。	2004/09/24
24:pickuplvc_sahen_wago		十二月に離婚し を通じて保田 っていたとこ 場にいた中村 県警に通報し	2004/09/19
27:pickuplvc_sahen_wago		の向かい側に 車を発進させ	2004/09/19
36:pickuplvc_sahen_wago	■副知事、三宅村の長谷川鴻村長ら約二十人が視察した。	島を三十日、 鴻村長ら約二十人が視察した。	2004/09/24
45:pickuplvc_sahen_wago	全国初の試みで、被害者からは「捜査が実感でき、不安の解消になる」と好評だ。	全国初の試みで、被害者からは「捜査が実感でき、不安が解消できる」と好評だ。	2004/10/01
102:pickuplvc_others_wago	威勢のいい店員の呼び込みが飛び交う中、大きな買い物袋を提げた主婦らが品定めする姿が見られた。	威勢のいい店員が呼び込む中、大きな買い物袋を提げた主婦らが品定めする姿が見られた。	2004/09/13
125:pickuplvc_sahen_wago	急速に高値修正が進み、ドバイは十二月中旬、一年四カ月ぶりに二〇ドルを割り込んだ。	急速に高値修正され、ドバイは十二月中旬、一年四カ月ぶりに二〇ドルを割り込んだ。	2004/09/13
126:pickuplvc_sahen_wago	原油高に連動して石油化学の基礎原料ナフサが高騰、石化メーカーはコスト高を転嫁しようと一斉に値上げに動いた。	原油高に連動して石油化学の基礎原料ナフサが高騰、石化メーカーはコスト高を転嫁しようと一斉に値上げした。	2004/09/13
131:pickuplvc_sahen_wago	食品包装フィルム、ペットボトルなどは逆に値下げを求められ、未転嫁のまま年末を迎えた。	食品包装フィルム、ペットボトルなどは逆に値を下げさせられ、未転嫁のまま年末を迎えた。	2004/09/14
133:pickuplvc_sahen_wago	選考にあたって重視したのは、日本や海外の企業にも影響を与えたり、目標と擬せられるようなモデルになり得たかどうかだ。	選考にあたって重視したのは、日本や海外の企業にも影響したり、目標と擬せられるようなモデルになり得たかどうかだ。	2004/09/13
134:pickuplvc_sahen_sahen	花王は、常に徹底的な変革に挑戦している。	花王は、常に徹底的に変革しようとしている。	2004/10/04
	日本経済新聞社が創設した「日経ミレニアム企業賞」の選考	日本経済新聞社が創設した「日経ミレニアム企業賞」を選	

Purpose of paraphrase corpus

- To analyze how we paraphrase
- To automate paraphrase generation
  - Induction of paraphrase patterns
  - Identification of useful linguistic knowledge
- As a standard test-set
  - Evaluation of paraphrase generation systems

Contributes to activate the research field

Contents

- Background
- Issues and our approach
- Constructed corpora
  - Work on grants for COE fellow
    "Creating large scale corpus with contextual information"
  - Discussion
- Conclusion and future direction

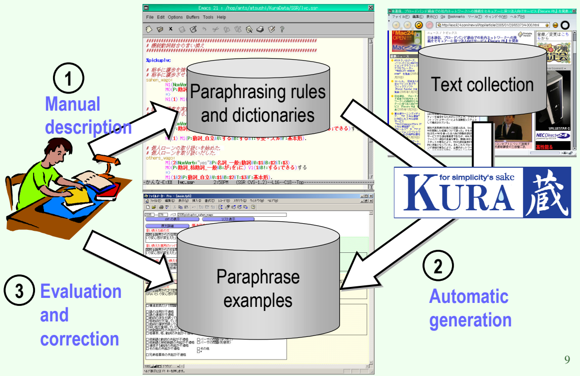
Issue

- How to effectively collect diverse examples?
  - What we should collect?
    - Hard to collect every possible paraphrase
    - Human annotators tend to produce biased examples [Shirai et al., 2001][Kinjo et al., 2003][Shimohata, 2004]
  - Frequencies of each paraphrase category are different
    - Paraphrase categories [Fujita et al., 2004]
      - "Passive sentences to active ones,"
      - "Transitive verb phrases to intransitive ones,"
      - "Division of complex sentences,"
      - etc.

Our approach & present goal

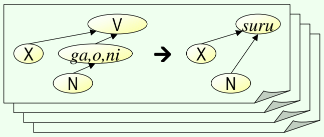
- How to effectively collect diverse examples?
  - Use an existing paraphrase generation system
    → reduces human labor and bias
  - Collect examples for each paraphrase category
    → enables us to collect every possible paraphrase
    - Capture the range of a paraphrase category using a set of paraphrasing rules
- Present goal
  - To confirm feasibility of the approach through creating corpora (pl. of corpus)

Corpus construction procedure


1 Manual description
Paraphrasing rules and dictionaries
Text collection
for simplicity's sake KURA 蔵
2 Automatic generation
Paraphrase examples
3 Evaluation and correction

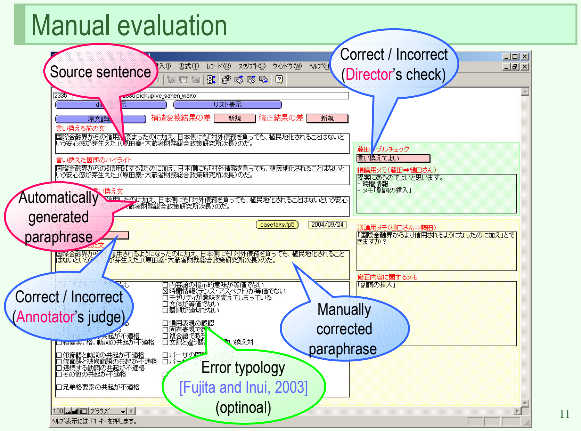
Paraphrasing rules

- To collect paraphrasable sentences
  - Pairs of dependency trees
  - Implemented on a paraphrase generation system



X: variable for any word
N: variable for a noun
V: variable for a verb
ga,o,ni,suru: words

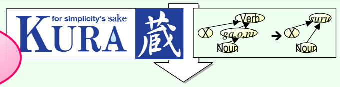
Manual evaluation


Source sentence
Correct / Incorrect (Director's check)
Automatically generated paraphrase
Correct / Incorrect (Annotator's judge)
Manually corrected paraphrase
Error typology [Fujita and Inui, 2003] (optinoal)

Example

Source sentence

DNAの二重らせん構造解明は、生命科学やバイオテクノロジーに劇的な進歩をもたらすことになった。



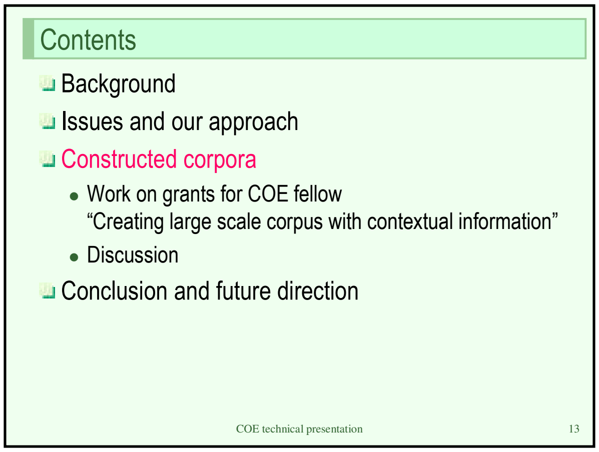
Automatically generated paraphrase

DNAの二重らせん構造解明は、生命科学やバイオテクノロジーに劇的な進歩することになった。

Manually corrected paraphrase

DNAの二重らせん構造解明は、生命科学やバイオテクノロジーを劇的に進歩させることになった。

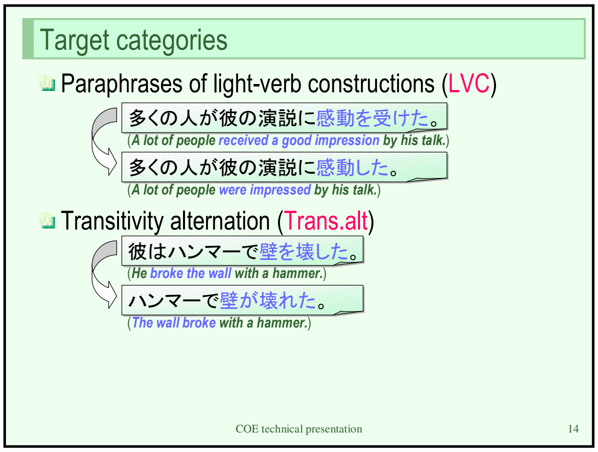
Re-assigning cases, modifying conjugation forms, changing voice, etc.



Contents

- Background
- Issues and our approach
- Constructed corpora
  - Work on grants for COE fellow
    "Creating large scale corpus with contextual information"
  - Discussion
- Conclusion and future direction

COE technical presentation 13



Target categories

- Paraphrases of light-verb constructions (LVC)

多くの人が彼の演説に感動を受けた。
(*A lot of people received a good impression by his talk.*)

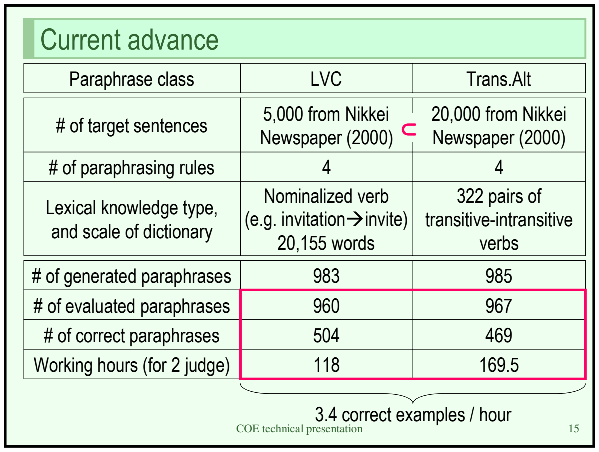
多くの人が彼の演説に感動した。
(*A lot of people were impressed by his talk.*)

- Transitivity alternation (Trans.alt)

彼はハンマーで壁を壊した。
(*He broke the wall with a hammer.*)

ハンマーで壁が壊れた。
(*The wall broke with a hammer.*)

COE technical presentation 14

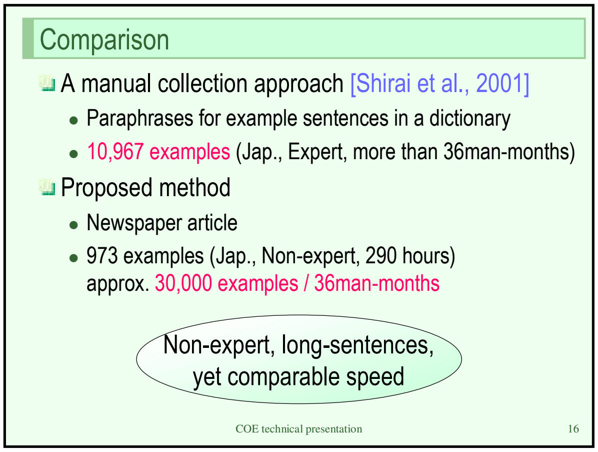


Current advance

Paraphrase class	LVC	Trans.Alt
# of target sentences	5,000 from Nikkei Newspaper (2000) ⊂	20,000 from Nikkei Newspaper (2000)
# of paraphrasing rules	4	4
Lexical knowledge type, and scale of dictionary	Nominalized verb (e.g. invitation→invite) 20,155 words	322 pairs of transitive-intransitive verbs
# of generated paraphrases	983	985
# of evaluated paraphrases	960	967
# of correct paraphrases	504	469
Working hours (for 2 judge)	118	169.5

3.4 correct examples / hour

COE technical presentation 15

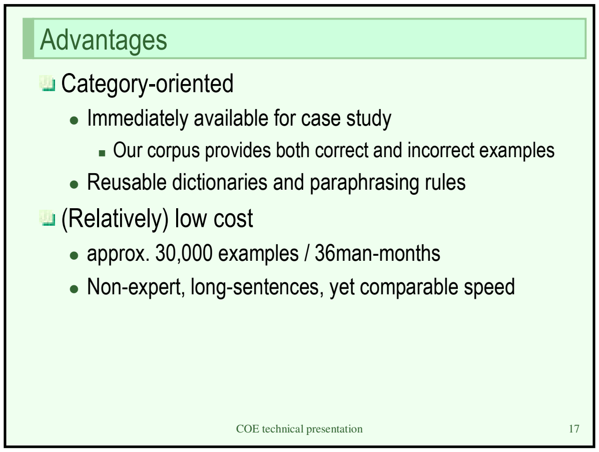


Comparison

- A manual collection approach [Shirai et al., 2001]
  - Paraphrases for example sentences in a dictionary
  - 10,967 examples (Jap., Expert, more than 36man-months)
- Proposed method
  - Newspaper article
  - 973 examples (Jap., Non-expert, 290 hours)
    approx. 30,000 examples / 36man-months

Non-expert, long-sentences,
yet comparable speed

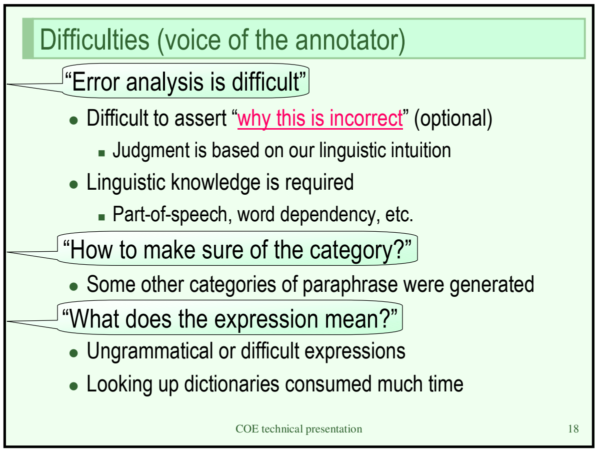
COE technical presentation 16



Advantages

- Category-oriented
  - Immediately available for case study
    - Our corpus provides both correct and incorrect examples
  - Reusable dictionaries and paraphrasing rules
- (Relatively) low cost
  - approx. 30,000 examples / 36man-months
  - Non-expert, long-sentences, yet comparable speed

COE technical presentation 17



Difficulties (voice of the annotator)

"Error analysis is difficult"

- Difficult to assert "why this is incorrect" (optional)
  - Judgment is based on our linguistic intuition
- Linguistic knowledge is required
  - Part-of-speech, word dependency, etc.

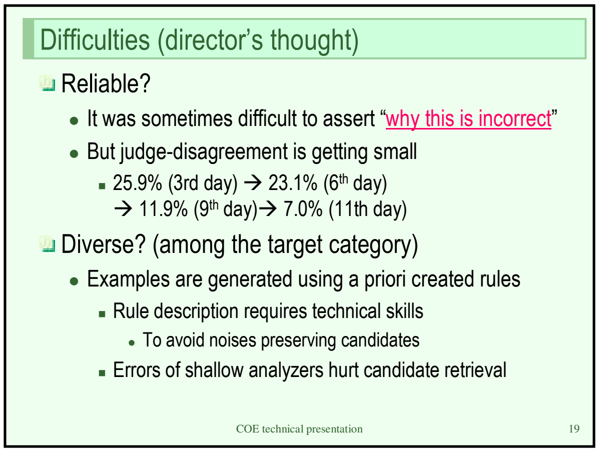
"How to make sure of the category?"

- Some other categories of paraphrase were generated

"What does the expression mean?"

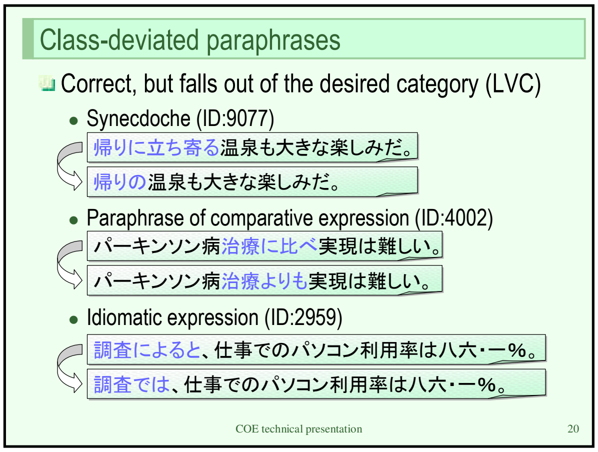
- Ungrammatical or difficult expressions
- Looking up dictionaries consumed much time

COE technical presentation 18



Difficulties (director's thought)

- Reliable?
  - It was sometimes difficult to assert "why this is incorrect"
  - But judge-disagreement is getting small
    - 25.9% (3rd day) → 23.1% (6th day) → 11.9% (9th day)→ 7.0% (11th day)
- Diverse? (among the target category)
  - Examples are generated using a priori created rules
    - Rule description requires technical skills
      - To avoid noises preserving candidates
    - Errors of shallow analyzers hurt candidate retrieval



Class-deviated paraphrases

- Correct, but falls out of the desired category (LVC)
  - Synecdoche (ID:9077)

帰りに立ち寄る温泉も大きな楽しみだ。

帰りの温泉も大きな楽しみだ。

  - Paraphrase of comparative expression (ID:4002)

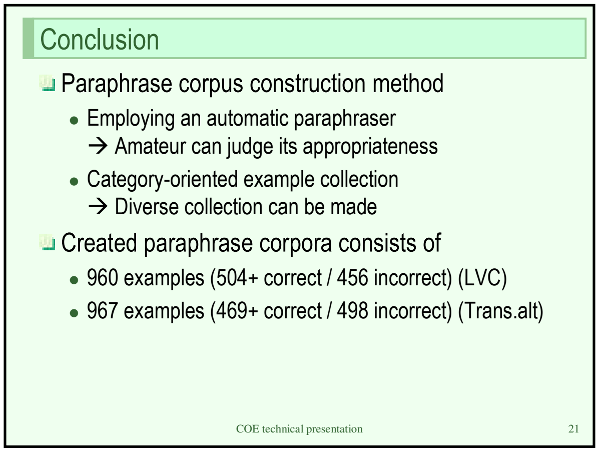
パーキンソン病治療に比べ実現は難しい。

パーキンソン病治療よりも実現は難しい。

  - Idiomatic expression (ID:2959)

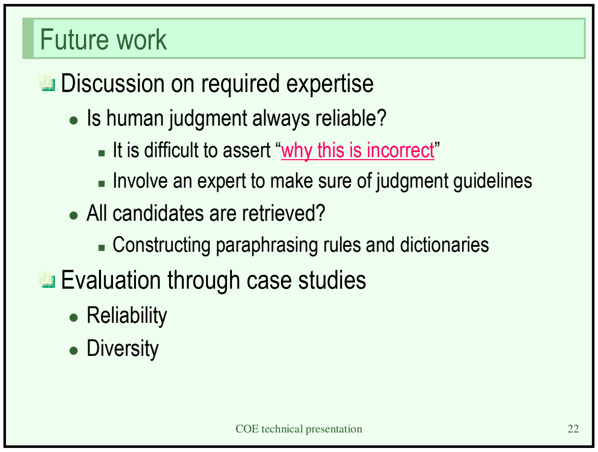
調査によると、仕事でのパソコン利用率は八六・一%。

調査では、仕事でのパソコン利用率は八六・一%。



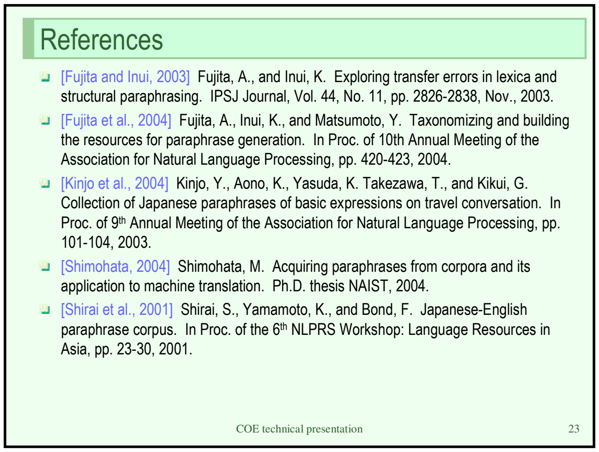
Conclusion

- Paraphrase corpus construction method
  - Employing an automatic paraphraser → Amateur can judge its appropriateness
  - Category-oriented example collection → Diverse collection can be made
- Created paraphrase corpora consists of
  - 960 examples (504+ correct / 456 incorrect) (LVC)
  - 967 examples (469+ correct / 498 incorrect) (Trans.alt)



Future work

- Discussion on required expertise
  - Is human judgment always reliable?
    - It is difficult to assert "why this is incorrect"
    - Involve an expert to make sure of judgment guidelines
  - All candidates are retrieved?
    - Constructing paraphrasing rules and dictionaries
- Evaluation through case studies
  - Reliability
  - Diversity



References

- [Fujita and Inui, 2003] Fujita, A., and Inui, K. Exploring transfer errors in lexica and structural paraphrasing. IPSJ Journal, Vol. 44, No. 11, pp. 2826-2838, Nov., 2003.
- [Fujita et al., 2004] Fujita, A., Inui, K., and Matsumoto, Y. Taxonomizing and building the resources for paraphrase generation. In Proc. of 10th Annual Meeting of the Association for Natural Language Processing, pp. 420-423, 2004.
- [Kinjo et al., 2004] Kinjo, Y., Aono, K., Yasuda, K. Takezawa, T., and Kikui, G. Collection of Japanese paraphrases of basic expressions on travel conversation. In Proc. of 9th Annual Meeting of the Association for Natural Language Processing, pp. 101-104, 2003.
- [Shimohata, 2004] Shimohata, M. Acquiring paraphrases from corpora and its application to machine translation. Ph.D. thesis NAIST, 2004.
- [Shirai et al., 2001] Shirai, S., Yamamoto, K., and Bond, F. Japanese-English paraphrase corpus. In Proc. of the 6th NLPRS Workshop: Language Resources in Asia, pp. 23-30, 2001.