日英対訳パターンの自動抽出に向けて

鳥取大学大学院工学研究科

道祖尾 太祐　　村上 仁一

徳久 雅人　　池原 悟

研究の背景

機械翻訳において
翻訳知識の獲得が重要な課題の一つ

↓

結合価文法が提案

結合価文法 … 用言を中心に単文レベルで翻訳

↓

用言や単文にこだわらない翻訳が必要

翻訳精度を向上させる方法

同じ意味を持つ“日本語表現”と“英語表現”を対にした“日英対訳パターン”の作成

(日英対訳パターンの例)

日本語表現	英語表現
「するために」	「in order to」
「*N*から*N*まで」	「from *N* to *N*」

- 日本語表現，英語表現を別々に抽出する方法は既に提案

問題点

- 日本語表現と英語表現の意味的対応を自動的に行い，日英対訳パターンとして抽出することは困難
- 大量の日英対訳パターンの作成は人手では困難

↓

人手での作成を補助する方法が必要

本研究の目的

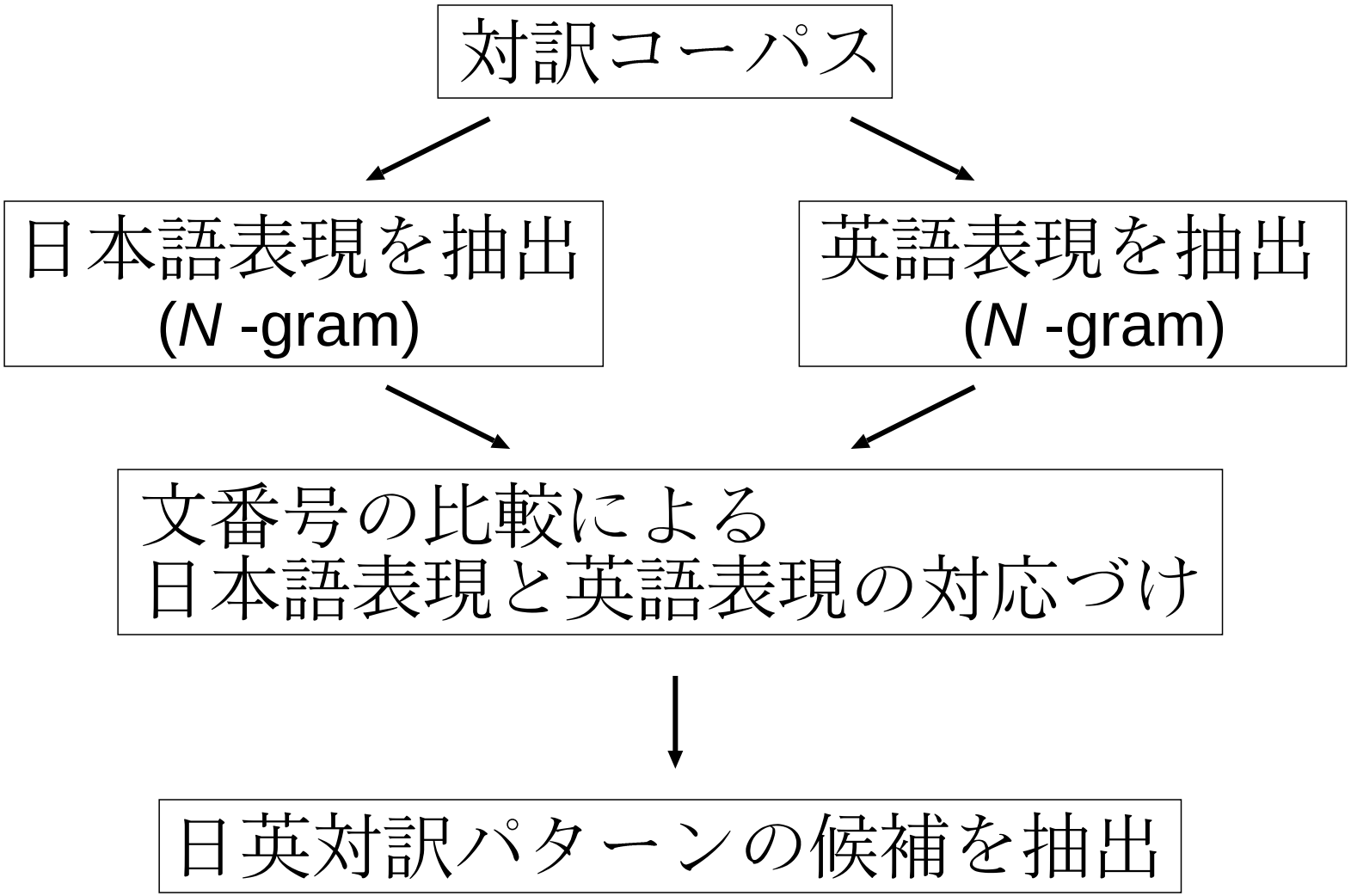
対訳コーパスから日英対訳パターンの候補を
自動的に抽出する方法の提案と，有効性の調査

対訳コーパス

日本文一文と英文一文が対応

文番号	日本文	英文
(1)	風呂が熱い.	the bath is too hot.
(2)	彼は立派な文を書く.	he writes a fine style.
(3)	これは別の品です.	this is a different article.

提案手法


対訳コーパス
日本語表現を抽出
(*N* -gram)
英語表現を抽出
(*N* -gram)
文番号の比較による
日本語表現と英語表現の対応づけ
日英対訳パターンの候補を抽出

日本語表現および英語表現の抽出方法

N -gram統計処理方法：池原, 白井, 河岡, “大規模コーパスからの連鎖型および離散型の共起表現の自動抽出法”, 1995

… 複数の文から共通の文字列を自動的に発見し，抽出する方法

- 連鎖型共起表現*N*-gram統計処理方法
  … 連続的な共通の文字列を抽出
  (例)「するために」
- 離散型共起表現*N*-gram統計処理方法
  … 離れた場所にある共通の文字列を抽出
  (例)「全く～ない」

連鎖型共起表現*N* -gram統計処理方法

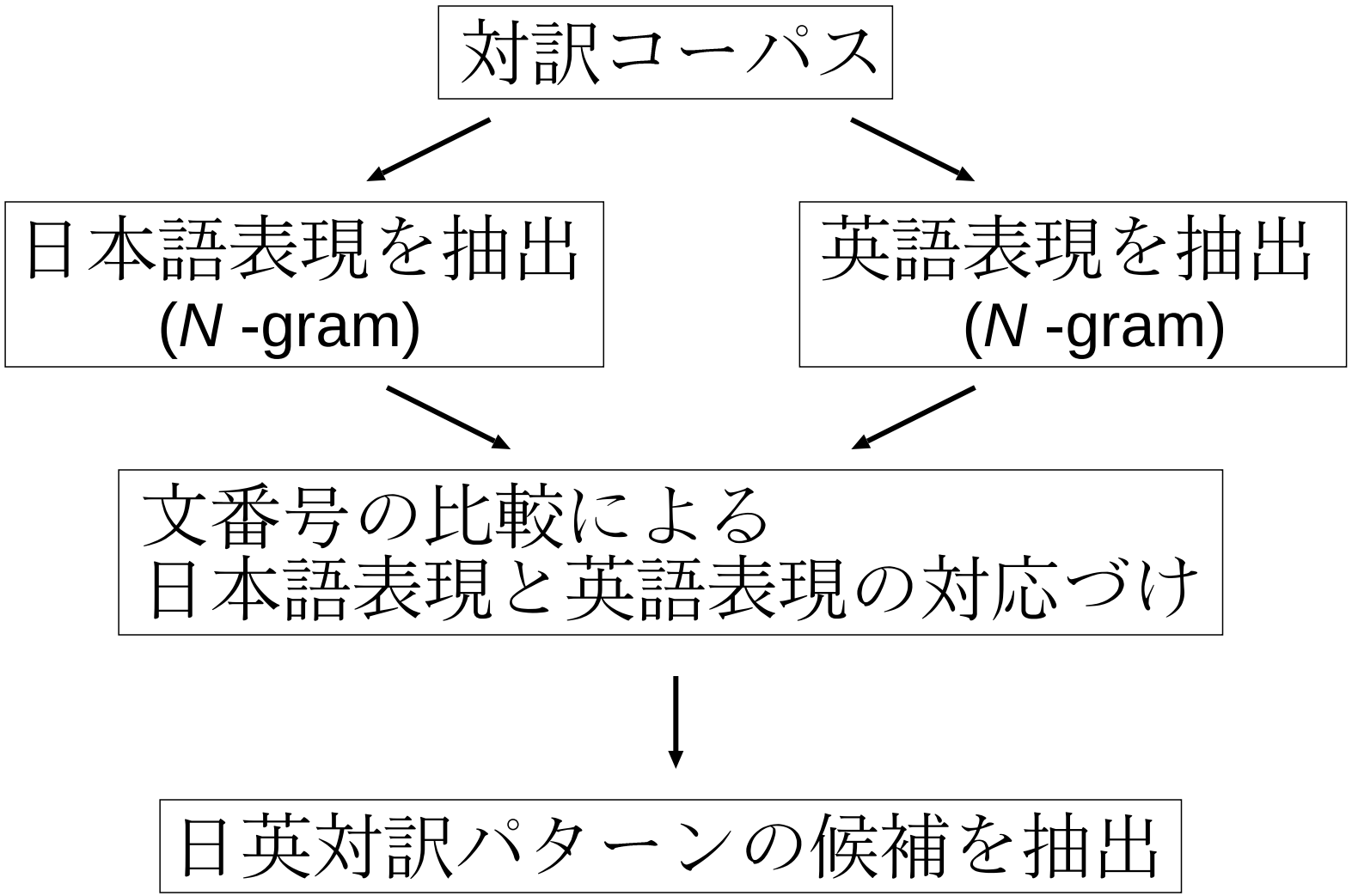
(例)　文(1). A ｜ＢＣＤ｜ＥＦ

文(2). G ｜ＢＣＤ｜H

文(3). ＩＪＢＣ

・無抑制型 … 「ＢＣＤ」，「ＢＣ」，「ＣＤ」が抽出
・強抑制型 … 「ＢＣＤ」が抽出
・弱抑制型 … 「ＢＣＤ」，「ＢＣ」が抽出

提案手法


対訳コーパス
日本語表現を抽出
(*N* -gram)
英語表現を抽出
(*N* -gram)
文番号の比較による
日本語表現と英語表現の対応づけ
日英対訳パターンの候補を抽出

日本語表現と英語表現の対応

対訳コーパスから抽出された日本語表現，
英語表現を含む文の文番号を比較

↓

同じ文番号の日本語表現と英語表現は
日英対訳パターンである可能性が高いと仮定

↓

日英対訳パターンの候補を自動的に抽出

日本語表現と英語表現を対応させる方法

(対訳コーパスの例)

文番号	日本文	英文
(1)	a b c d	A B C D W X
(2)	e b c f	E B C F Y Z

文番号の一致率

$$\text{文番号の一致率} = \frac{\text{文番号の一致数}}{\text{表現の抽出回数}}$$

(文番号)

日本語表現	英語表現
「b c」… 文番号(1), (2)	「B C」… 文番号(1), (2)
	「W X」… 文番号(1)
	「Y Z」… 文番号(2)

(文番号の一致率)

日	英	一致率
「b c」	「B C」	100%(文番号(1), (2)のうち，(1), (2)が一致)
「b c」	「W X」	50%(文番号(1), (2)のうち，(1)が一致)
「b c」	「Y Z」	50%(文番号(1), (2)のうち，(2)が一致)

閾値を設定し，日英対訳パターンの候補を抽出

評価実験

実験の目的

日英対訳パターンの候補を自動的に
抽出する方法の有効性の調査

↓

連鎖型共起表現に対して実験

実験の手順

対訳コーパス：対訳コーパスの種類
単語単位，名詞の置換

日本語表現を抽出
(*N* -gram : 連鎖型，離散型)

英語表現を抽出
(*N* -gram : 連鎖型，離散型)

文番号の比較による
日本語表現と英語表現の対応づけ

日英対訳パターンの候補を抽出：閾値の設定

評価(人手)

実験条件

1. 日本語表現と英語表現の単位

・単語単位

… 単語を単位として文字列を抽出することで，
意味を持たない文字列の削除が可能

・名詞の置換

… 名詞を置換することで，意味として
まとまりを持つ表現の抽出が可能

名詞の置換により，表現の意味や構文形成に
重要な文字列を見失う可能性

→ 二つの方法で実験

2. 実験に用いる対訳コーパス

複数の対訳辞書から抽出した単文を使用

- 単語単位の場合 … 8,500文
- 名詞を置換した場合 … 28,000文

3. 連鎖型共起表現の抑制方法

日本語表現および英語表現は，連鎖型共起表現
N -gram統計処理方法の強抑制型と弱抑制型で抽出

4. 日英対訳パターンの候補の抽出(閾値の設定)

・文番号の一致数が3以上の日英対訳パターン
かつ，
・文番号の一致率が50%以上の日英対訳パターン

評価方法

抽出された日英対訳パターンの候補の
上位50個を人手で評価し，正解率を算出

＜評価の分類＞

○：完全に対訳であると判断されるもの
△：ほぼ対訳であると判断されるもの
×：対訳ではないと判断されるもの

＜正解率＞

$$\text{正解率(1)} = \frac{\text{○の数}}{\text{評価対象の総数(50)}}$$

$$\text{正解率(2)} = \frac{\text{○と△の数}}{\text{評価対象の総数(50)}}$$

実験結果

1. 単語単位の場合の強抑制型

2. 単語単位の場合の弱抑制型

3. 名詞を置換した場合の強抑制型

4. 名詞を置換した場合の弱抑制型

1. 単語単位の場合の強抑制型(抽出数74)

評価	日	英
“○” (22／50)	彼はまだ これは私の	he is still this is my
“△” (10／50)	は丘の上にある と結婚した	on the hill she married a
“×” (18／50)	の天才だ へ行っている	he is a the ship is

2. 単語単位の場合の弱抑制型(抽出数161)

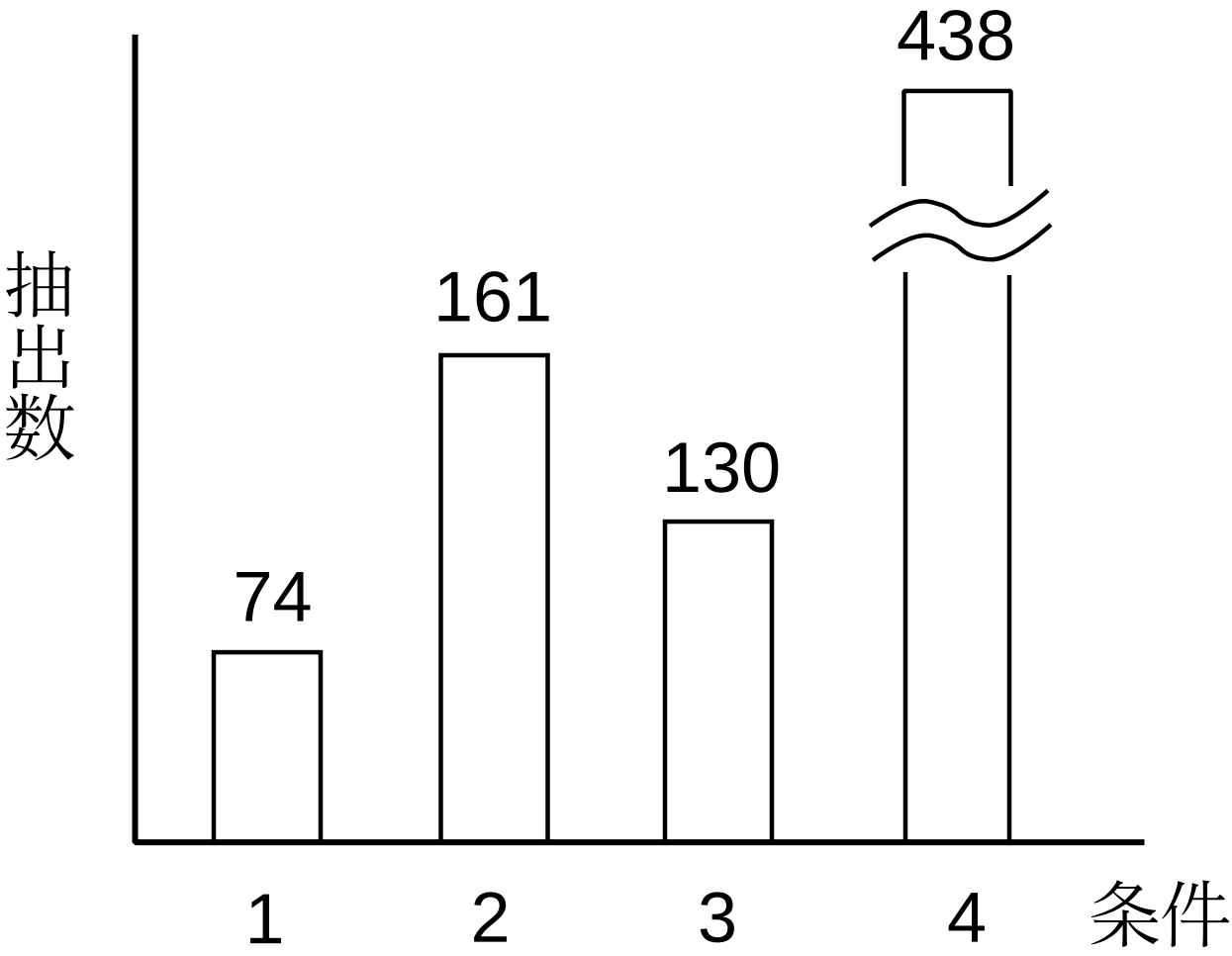
評価	日	英
“○” (12／50)	彼はまだ これは私の	he is still this is my
“△” (17／50)	で学校へ と言っている	to school by bus they complain of the
“×” (21／50)	の天才だ 立てられぬ	he is a people will talk

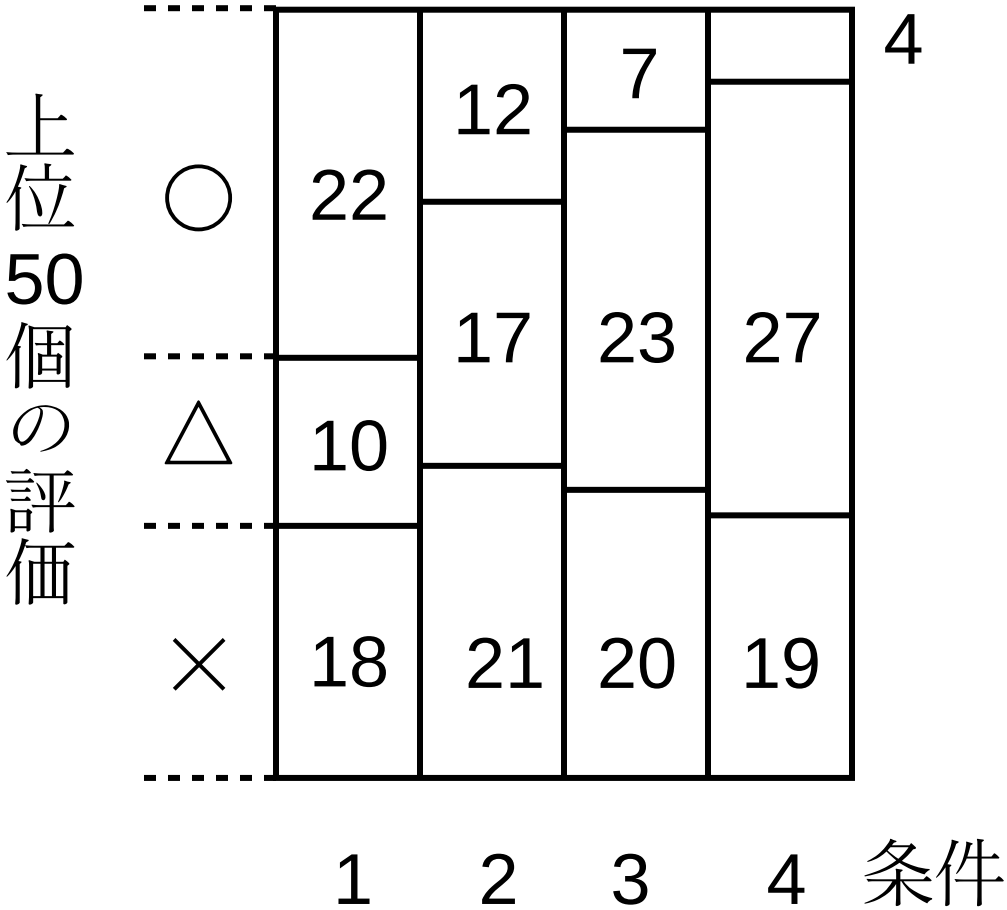
3. 名詞を置換した場合の強抑制型(抽出数130)

評価	日	英
“○” (7／50)	*N* は*N* が短い *N* と*N* は	*N* is short in *N* *N* and *N*
“△” (23／50)	*N* をもっている *N* を*N* に行った	*N* has a *N* went to
“×” (20／50)	この*N* は*N* が *N* は*N* に甘い	*N* of *N* to *N* ’s *N*

4. 名詞を置換した場合の弱抑制型(抽出数438)

評価	日	英
“○” (4／50)	*N* と *N* は *N* が安い	*N* and *N* *N* is low
“△” (27／50)	*N* を買った *N* をもっている	*N* bought a *N* *N* has a
“×” (19／50)	なかなかの*N* だ *N* に立っていた	*N* is a *N* of the *N*


438
抽出数
161
130
74
1
2
3
4
条件


上位50個の評価
○
△
×
22
12
7
4
10
17
23
27
18
21
20
19
1
2
3
4
条件

正解率

	正解率(1)	正解率(2)
1. 単語単位 (強抑制型)	44% (22／50)	64% (32／50)
2. 単語単位 (弱抑制型)	24% (12／50)	58% (29／50)
3. 名詞を置換 (強抑制型)	14% (7／50)	60% (30／50)
4. 名詞を置換 (弱抑制型)	8% (4／50)	62% (31／50)

考察1

単語単位の場合，名詞を置換した場合に
対して，それぞれ強抑制型，弱抑制型で実験

＜正解率＞

正解率(1)：単語単位の強抑制型で一番高い値

正解率(2)：平均61%

→ 日英対訳パターンの候補を自動的に
抽出できる見通し

＜表現の単位と抑制型＞

条件の選択は今後の検討が必要

考察2

抽出された日英対訳パターン
… ほぼ対訳(評価△)：多

↓

人手での修正により，完全な日英対訳パターンの収集が可能

＜修正の例＞… 下線部分が修正箇所

	日	英
(修正後)	この*N* では*N*	in this *N*
	この*N* では*N*	*N* in this *N*
(修正後)	*N* を買った	*N* bought a *N*
	N は*N* を買った	*N* bought a *N*

人手での修正が必要であるが,
日英対訳パターンの作成を補助

まとめ

日英対訳パターンの候補を自動的に抽出する方法を提案

↓

・正解率(1) … 単語単位の強抑制型で一番高い値
・正解率(2) … 平均61%

↓

本手法の有効性を確認

今後の課題

・離散型共起表現の抽出
・単文の他，重文や複文からの表現の抽出
・閾値の設定に関する調査