概要

機械翻訳において，熟語や連語のような意味的にまとまりを持つ表現を単位として翻訳を行う方法が注目されている．この場合，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” を大量に作成する必要がある．この日英対訳パターンの意味的対応は人手で判断するため，大量の日英対訳パターンを人手で作成することは困難である．そこで，人手での作成を補助するため，日英対訳パターンの候補を自動抽出する方法が必要となる．

そこで本研究では，熟語や連語のような意味的にまとまりを持つ表現を対象とした日英対訳パターンを作成するため，N-gram 統計処理方法により，原文に 2 回以上出現する日本語表現および英語表現を抽出した．そして，日本語表現および英語表現を含む文の文番号を比較し，その相互関係を用いることで日本語表現と英語表現を自動的に対応付けた．

上記の手法を単文 10,000 文，50,000 文，80,000 文に適用し，抽出した日英対訳パターンの候補を人手で評価した．その結果，約 84%の候補が人手による判断で日英対訳パターンを作成できることが分かった．

目 次

図 目 次

表 目 次

1　はじめに

機械翻訳において，熟語や連語のような意味的にまとまりを持つ表現を単位として翻訳を行う方法が注目されている．この場合,「するために」と「in order to」のように同じ意味を持つ “日本語表現” と “英語表現” を対にした “日英対訳パターン” を大量に作成する必要がある．この日英対訳パターンの意味的対応は人手で判断するため，大量の日英対訳パターンを人手で作成することは困難である．そこで，人手での作成を補助するため，日英対訳パターンの候補を自動抽出する方法が必要となる．

従来，対訳コーパスから日英対訳パターンの候補を自動抽出する方法が提案されてきた [1][2]．しかし，自動抽出された日英対訳パターンは，単語や合成名詞が多かった．

そこで本研究では，熟語や連語のような意味的にまとまりを持つ表現を対象とした日英対訳パターンの作成を目指し，対訳コーパスから日英対訳パターンの候補を自動抽出する方法を提案する．そして，自動抽出した日英対訳パターンの候補を人手で評価し，今後，日英対訳パターンを作成する際に有効な指針を示す．その結果，約 84%の候補が人手による判断で日英対訳パターンを作成できることが分かった．また，意味的にまとまりを持つ表現を対象とした日英対訳パターンを作成することができた．

以下，2 章では従来の手法について，3 章では日英対訳パターンの候補を自動抽出する手順について，4 章では実験について，5 章では実験結果について，そして 6 章では考察について，最後に 7 章で結論を述べる．

2　従来の手法

従来，対訳コーパスから日英対訳パターンの候補を自動抽出する方法が提案されてきた．熊野ら [1] は，言語的な情報と統計的な情報を利用し日本語表現と英語表現の対応関係を推定した．熊野らの方式は，言語間で対応付けが行える単位をユニットと定義し，日本語ユニット中の各内容語に対して機械翻訳対訳辞書から得られる訳語候補群と，英語ユニット中の内容語との対応度である対応関係確信度を求め，対応関係を推定している．

文献 [1] から引用し，説明する．

1. 日本語ユニット中の内容語リスト J を作成 $\cdots$ J $= \{J_1, J_2, \cdots, J_m\}$　(語数 m)
2. 英語ユニット中の内容語リスト E を作成 $\cdots$ E $= \{E_1, E_2, \cdots, E_n\}$　(語数 n)
3. J 中の各 J_i に対して日英対訳辞書を参照し，$J_i \rightarrow E_j$ の関係にある E_j を全て選定 ($J_i \rightarrow E_j$ は，J_i の訳語候補に E_j を含んでいることを示している)
4. J 中の各 J_i のうち，いずれかの E_j に対して $J_i \rightarrow E_j$ の関係にあるものの語数を x とする
5. E 中の各 E_j のうち，いずれかの J_i に対して $J_i \rightarrow E_j$ の関係にあるものの語数を y とする
6. 対応関係確信度 $\cdots$ P(JU, EU) $=$ (x + y) / (m + n)

結果を文献 [1] から引用し，例 1 に紹介する．言語的な情報と統計的な情報を利用することで，言語的な情報だけでは抽出できなかった未知語を抽出した．

(例 1)　合成名詞

最小加工寸法	minimum featuring size
素子分離領域	element separation region
オープンビット線方式	open bit line configuration
カラムアドレスストローブ	column address strobe
セルアレイ	cell array

未知語

ポリッシング	polishing
コレクタ	collector
積層する	to form

北村ら [2] は，日本語表現と英語表現の対応関係の推定に利用される相互情報量と Dice 係数を比較し，Dice 係数の方が優れていることを示した．そして，対訳コーパスに複数回出現する任意の長さの単語列に対して Dice 係数 (式 (a)) を基にした式 (式 (b)) を定義し対応関係を推定している．

$$Dice(X,Y) = \frac{2 \cdot f_{xy}}{f_x + f_y} \cdots \text{(a)}$$

$$\mathrm{sim}(w_J, w_E) = (\log_2 f_{je}) \cdot \frac{2 \cdot f_{je}}{f_j + f_e} \cdots \text{(b)}$$

w_J：日本語単語列　　w_E：英語単語列　　f_j：w_J の出現回数

f_e：w_E の出現回数　　f_{je}：w_J, w_E の同時出現回数

結果を文献 [2] から引用し，例 2 に紹介する．例 2 は抽出した対訳表現の上位の例である．実験では,「技術情報：technical information」のように 2 語以上で構成される対訳表現も抽出した.

(例 2)

会社	company
ライセンシー	licensee
販売店	distributor
契約 品	product
売り手	seller

例 1 や例 2 に示すように，従来の日英対訳パターンは単語や合成名詞が中心となっていた．そこで，本研究では，熟語や連語のような意味的にまとまりを持つ表現を対象とした日英対訳パターンの作成を目指す.

3　日英対訳パターンの候補の自動抽出手順

日英対訳パターンの候補を自動抽出する手順を以下に示す.

1. 日本語表現および英語表現を抽出
2. 日本語表現と英語表現の対応付け

それぞれについて説明する.

3.1　日本語表現および英語表現の自動抽出

本研究では，対訳コーパスから日本語表現および英語表現を抽出する場合，N-gram 統計処理方法 [3] を用いる．N-gram 統計処理方法は，二言語間から同時に表現を抽出することが困難であるため，まず日本語表現を抽出し，次に英語表現を抽出する．

N-gram 統計処理方法には，対象とする文字列に対して以下の 2 つの方法がある．

- 連鎖型共起表現 N-gram 統計処理方法 [3]
- 離散型共起表現 N-gram 統計処理方法 [3]

連鎖型共起表現とは,「するために」のように連続した文字列のことである．離散型共起表現とは,「もし〜ならば」のように離れた場所にある複数の文字列で構成される文字列のことである.

3.1.1 連鎖型共起表現 N-gram 統計処理方法

連鎖型共起表現 N-gram 統計処理方法は，複数の文から連続的な共通の文字列を抽出する方法である．図 1 を用いて説明する．図 1 の文 (1) は文字列「abc」，文 (2) は文字列「abd」である．図 1 からは「ab」が抽出される．

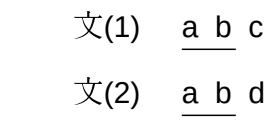

図 1: 連鎖型共起表現の抽出

連鎖型共起表現 N-gram 統計処理方法は，抽出の抑制について以下の 3 つの方法がある [4].

- 無抑制型
- 強抑制型
- 弱抑制型

無抑制型は，連続的な共通の文字列を全て抽出する方法である．強抑制型は，連続的な共通の文字列のうち，他の文字列と重なっている文字列は抽出しない方法である．弱抑制型は，他の文字列と重なっている文字列を抽出せず，他の文字列と重なっていても他の場所で独立している文字列を抽出する方法である．

3.1.2 離散型共起表現 N-gram 統計処理方法

離散型共起表現 N-gram 統計処理方法は，複数の文から離れた場所にある共通の文字列を抽出する方法である．図 2 を用いて説明する．図 2 の文 (1) は文字列「efghi」，文 (2) は文字列「efjhi」である．図 2 からは「ef〜hi」が抽出される．

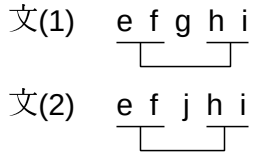

図 2: 離散型共起表現の抽出

離散型共起表現 N-gram 統計処理方法は，抽出の抑制について以下の 3 つの方法がある [4].

- 無抑制型
- 強抑制型
- 弱抑制型

無抑制型は，共起した文字列の全てを抽出する方法である．弱抑制型は，互いに異なる文字列のみを抽出する方法である．強抑制型は，弱抑制に加え，先頭と末尾の文字列間には，着目する文字列が 2 回以上出現しないものを抽出する方法である [4][7].

3.1.3 連鎖型共起表現と離散型共起表現

離散型共起表現は，離れた場所にある複数の文字列で構成される文字列で，連鎖型共起表現が共起したものと考えることができる．そこで，本研究では連鎖型共起表現を対象に日本語表現および英語表現を抽出する．次節で連鎖型共起表現 N-gram 統計処理方法について詳しく説明する．

3.1.4 連鎖型共起表現の抑制方法

連鎖型共起表現 N-gram 統計処理方法の無抑制型，強抑制型，弱抑制型について，図 3 を用いて説明する．図 3 の文 (1) は文字列「KLMNO」，文 (2) は文字列「PLMNQ」，文 (3) は文字列「LMR」で,「LMN」,「LM」,「MN」が連続的な共通の文字列である．

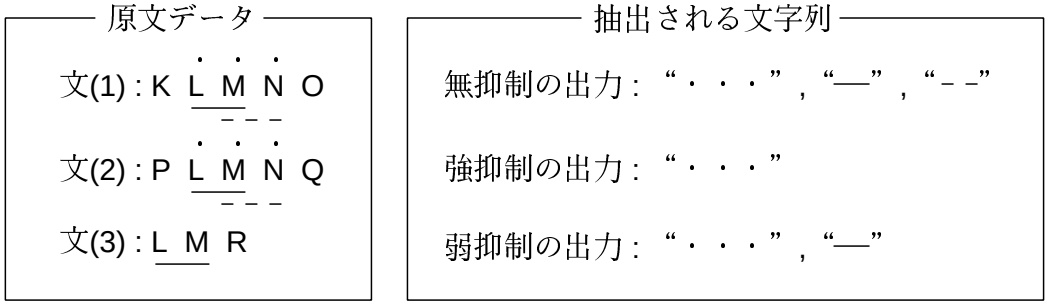

図 3: 連鎖型共起表現の抑制方法の例

無抑制型は，連続的な共通の文字列を全て抽出する．図 3 では,「LMN」,「LM」,「MN」が抽出される．

強抑制型は，連続的な共通の文字列のうち，他の文字列と重なっている文字列は抽出しない．図 3 では,「LMN」が抽出される.「LM」,「MN」は「LMN」と重なっているため抽出されない．

弱抑制型は，他の文字列と重なっている文字列を抽出せず，他の文字列と重なっていても他の場所で独立している文字列を抽出する．図 3 では,「LMN」,「LM」が抽出される.「LM」は「LMN」と重なっているが，文 (3) で独立しているため抽出される.「MN」は「LMN」と重なっていて，他の場所で独立していないため抽出されない．

3.2 日本語表現と英語表現の自動的な対応付け

3.2.1 文番号の一致率

本研究では，日本語表現および英語表現の抽出は連鎖型共起表現 N-gram 統計処理方法を用いる．しかし，N-gram 統計処理方法は二言語間から同時に表現を抽出することは困難であるため，別々に抽出した日本語表現と英語表現を対応付ける必要がある．そこで，日本語表現および英語表現が抽出された文の相互関係を用いて日本語表現と英語表現を意味的に対応付ける．

具体的には，日本語表現を含む文の文番号と英語表現を含む文の文番号を比較する．そして，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し，文番号が一致している割合が高い表現同士を日英対訳パターンの候補として自動抽出する [5]．本研究では，文番号が一致している割合を “文番号の一致率” と定義し，式 (1) で求める．

文番号の一致率 ＝ 文番号の一致数 ／ 抽出回数 … (1)

抽出回数とは，対訳コーパスから抽出された表現の抽出回数である．

3.2.2 日本語表現と英語表現を対応付ける手順

図 4 において，日英対訳パターンの候補を抽出する方法を説明する．図 4 では，日本語表現「bc」を含む日本文と，英語表現「BC」,「EF」を含む英文が対訳である (文番号 (1))．また，日本語表現「bc」を含む日本文と，英語表現「BC」,「IJ」を含む英文が対訳である (文番号 (2))．

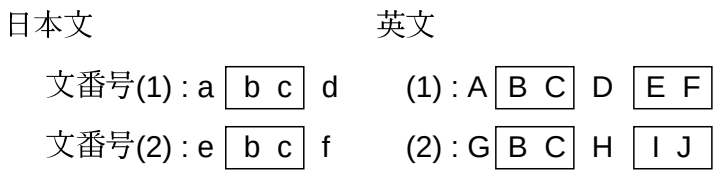

図 4: 対訳コーパスの例 (1)

図 4 の日本語表現と英語表現の文番号を表 1 に，文番号の一致率を表 2 に示す．

表 1: 日本語表現と英語表現の文番号

日本語表現	英語表現
「bc」 … 文番号 (1), (2)	「BC」 … 文番号 (1), (2)
	「EF」 … 文番号 (1)
	「IJ」 … 文番号 (2)

表 2: 文番号の一致率

日本語表現	英語表現	文番号の一致率
「bc」	「BC」	100%(文 (1), (2) のうち，(1), (2) が一致)
「bc」	「EF」	50%(文 (1), (2) のうち，(1) が一致)
「bc」	「IJ」	50%(文 (1), (2) のうち，(2) が一致)

そして，文番号の一致率が高い日本語表現「bc」と英語表現「BC」を日英対訳パターンの候補として抽出する．

4　実験

4.1　実験の目的

本研究では，人手での日英対訳パターンの作成を補助するため，文番号の一致率を求め，文番号の一致率が高い表現同士を日英対訳パターンの候補として自動抽出する．そして，抽出した日英対訳パターンの候補に対し，人手で評価を行う．

4.2　実験の手順

本実験の実験手順を図 5 に示す．

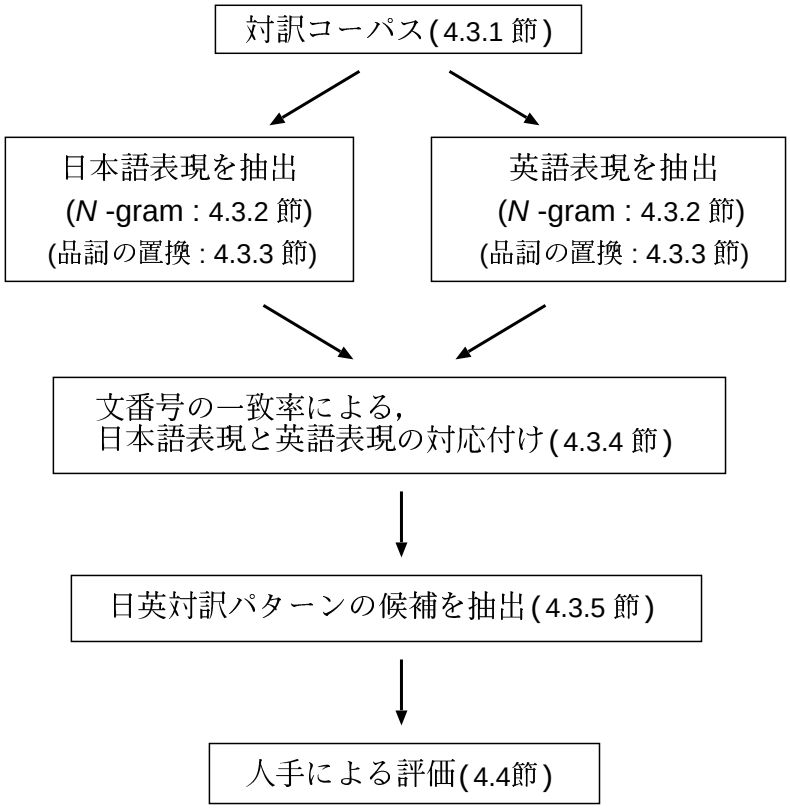

図 5: 実験手順

それぞれの処理について説明する．

4.3 実験条件

4.3.1 対訳コーパス

複数の対訳辞書 [6] から抽出した単文 10,000 文, 50,000 文, 80,000 文を対訳コーパスとして使用する．対訳コーパスの例を表 3 に示す．

表 3: 対訳コーパスの例 (2)

文番号	日本文	英文
1	トムとベティーは同い年だ．	Tom and Betty are of an age.
2	犬は忠実な動物である．	A dog is a faithful animal.
3	彼は英語を少々知っている．	He has a knowledge of English.
4	ピカソの絵を買った．	I bought a Picasso.
5	彼女はスペイン語がわかる．	She understands Spanish.

4.3.2 日本語表現および英語表現の抽出

日本語表現および英語表現を抽出する場合，強抑制型は意味を持たない文字列を抽出する可能性が低いため，本研究では連鎖型共起表現 N-gram 統計処理方法の強抑制型を用いる．N-gram 統計処理方法は二言語間から同時に表現を抽出することが困難であるため，まず，対訳コーパスの日本文から 2 回以上出現する日本語表現を抽出し，次に，対訳コーパスの英文から 2 回以上出現する英語表現を抽出する．

4.3.3 品詞の置換

日本語表現および英語表現を抽出する場合，意味的にまとまりを持つ表現を抽出するため，品詞ごとに単語を置換する方法 [7] が提案されている．本研究では，様々な種類の日英対訳パターンの候補を抽出するため，以下の 4 つの場合に対して実験を行う．

1. 文字を単位とした字面の場合
2. 名詞を置換した場合
3. 名詞・動詞を置換した場合
4. 名詞・動詞・副詞を置換した場合

それぞれについて説明する．

(1) 文字を単位とした字面の場合

図 6 を用いて説明する．図 6 の原文データは，文 (1) が「私は走る.」，文 (2) が「あなたは走る.」である．図 6 からは，連鎖型共起表現 N-gram 統計処理方法により，連続的な共通の文字列である「は走る.」が抽出される．

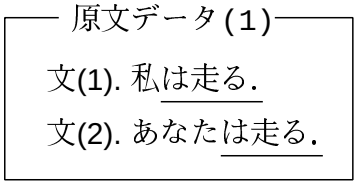

図 6: 連鎖型共起表現の抽出例 (字面)

(2) 名詞を置換した場合

図 6 の原文データの名詞を N に置換したデータを図 7 に示す．文 (1) が「N は走る.」，文 (2) が「N は走る.」である．図 7 からは，連鎖型共起表現 N-gram 統計処理方法により，連続的な共通の文字列である「N は走る.」が抽出される．

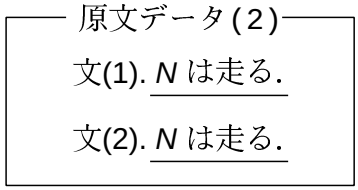

図 7: 連鎖型共起表現の抽出例 (名詞の置換)

(3) 名詞・動詞を置換した場合

図 8 を用いて説明する．図 8 の原文データは，文 (1) が「N を見る.」，文 (2) が「N を読む.」である．動詞を V に置換すると，文 (1) が「N を V.」，文 (2) が「N を V.」となり．連鎖型共起表現 N-gram 統計処理方法により，連続的な共通の文字列である「N を V.」が抽出される．

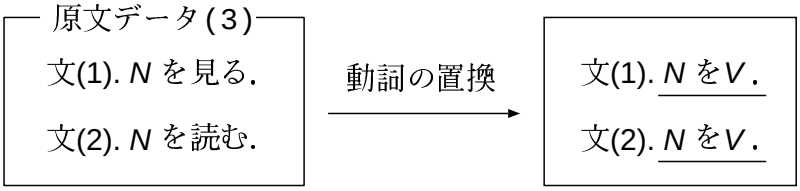

図 8: 連鎖型共起表現の抽出例 (名詞・動詞の置換)

(4) 名詞・動詞・副詞を置換した場合

図 9 を用いて説明する．図 9 の原文データは，文 (1) が「N をはやく V.」，文 (2) が「N を急いで V.」である．副詞を Adv に置換すると，文 (1) が「N を Adv V.」，文 (2) が「N を Adv V.」となり．連鎖型共起表現 N-gram 統計処理方法により，連続的な共通の文字列である「N を Adv V.」が抽出される．

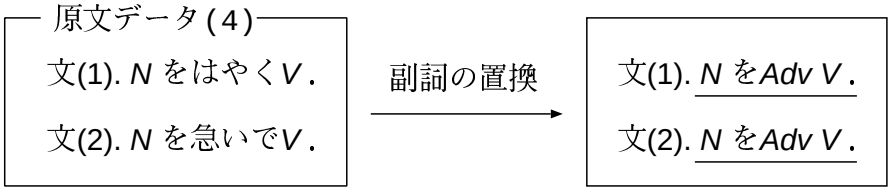

図 9: 連鎖型共起表現の抽出例 (名詞・動詞・副詞の置換)

4.3.4 日本語表現と英語表現の対応付け

対訳コーパスから別々に抽出した日本語表現と英語表現を対応付けるため，3.2.1 節で定義した文番号の一致率 (式 (1)) を求め，文番号の一致率が高い表現同士を日英対訳パターンの候補とする．

4.3.5 日英対訳パターンの候補の抽出

日英対訳パターンである可能性が高いものを効率良く抽出するため，文番号の一致数が 2 以上で，かつ，文番号の一致率が 50%以上のものを日英対訳パターンの候補として抽出する．

- 文番号の一致数が 2 以上としたのは，文番号の一致数が 1 の場合は，偶然文番号が一致している可能性があると判断したためである．
- 文番号の一致率が 50%以上としたのは，文番号の一致率が 50%より低い場合は，日英対訳パターンである可能性が低いと判断したためである．

4.4 評価方法

日英対訳パターンの候補を自動抽出した後，人手で評価を行い，日英対訳パターンを決定する．評価は，抽出した日英対訳パターンの候補の日本語表現に対して英語表現が意味的に対応しているかを判断し，三つに分類する．

○：完全に対訳であると判断される日英対訳パターン

△：ほぼ対訳であると判断される日英対訳パターン

×：対訳ではないと判断される日英対訳パターン

評価 “△” には，人手で修正を行うことで，日英対訳パターンが作成できる可能性があるものも含むこととする．

そして，正解率を式 (2) で求める．

正解率 ＝ “○” と “△” の数 ／ 評価対象の総数 … (2)

5　実験結果

5.1　日英対訳パターンの候補数

本研究では，複数の対訳辞書から抽出した単文を用いて実験を行った．連鎖型共起表現を対象にした実験で抽出した日英対訳パターンの候補数を表 4 に示す．

表 4: 日英対訳パターンの候補数

	10,000 文	50,000 文	80,000 文
文字を単位とした字面	112	1,317	2,735
名詞の置換	133	1,058	2,181
名詞・動詞の置換	113	672	1,296
名詞・動詞・副詞の置換	108	613	1,198

10,000 文ではほぼ 110 個の日英対訳パターンの候補を抽出した．50,000 文と 80,000 文では文字を単位とした字面の場合で抽出した日英対訳パターンの候補が多く，品詞の置換を行うと日英対訳パターンの候補が少なくなった．日英対訳パターンの候補の分布を図 10 に示す．横軸が対訳コーパスの文数，縦軸が実験で抽出した日英対訳パターンの候補数である．実験は 10,000 文，50,000 文，80,000 文に対して以下の条件で行った．

1. 文字を単位とした字面の場合
2. 名詞を置換した場合
3. 名詞・動詞を置換した場合
4. 名詞・動詞・副詞を置換した場合

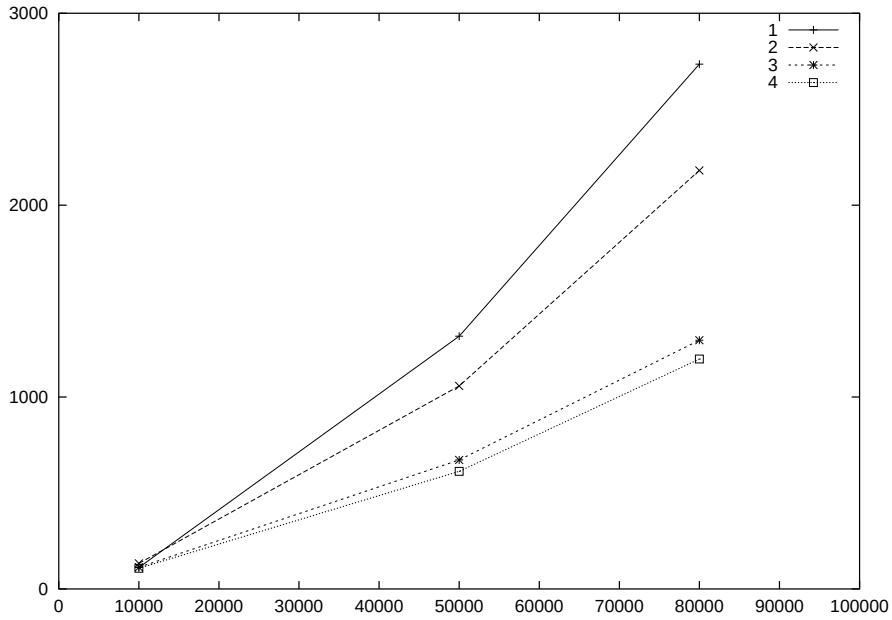

図 10: 日英対訳パターンの候補の分布

文数を増やせば日英対訳パターンの候補数が急激に増えていることが分かった.

全体では，80,000 文の文字を単位とした字面の場合で抽出した日英対訳パターンの候補が多かった．そこで，80,000 文の文字を単位とした字面の場合について詳しく述べる．また，品詞の置換を行った場合の結果を探るため，80,000 文の名詞・動詞・副詞を置換した場合についても詳しく述べる．他の条件の場合は付録参照.

5.2 80,000 文の字面の結果

5.2.1 日本語表現および英語表現の抽出

実験では，まず対訳コーパスの日本文から日本語表現を 13,629 個抽出し，次に英文から英語表現を 12,691 個抽出した．日本語表現の抽出回数の分布を図 11 に，英語表現の抽出回数の分布を図 12 に示す．縦軸が表現の出現回数，横軸が表現の個数である.

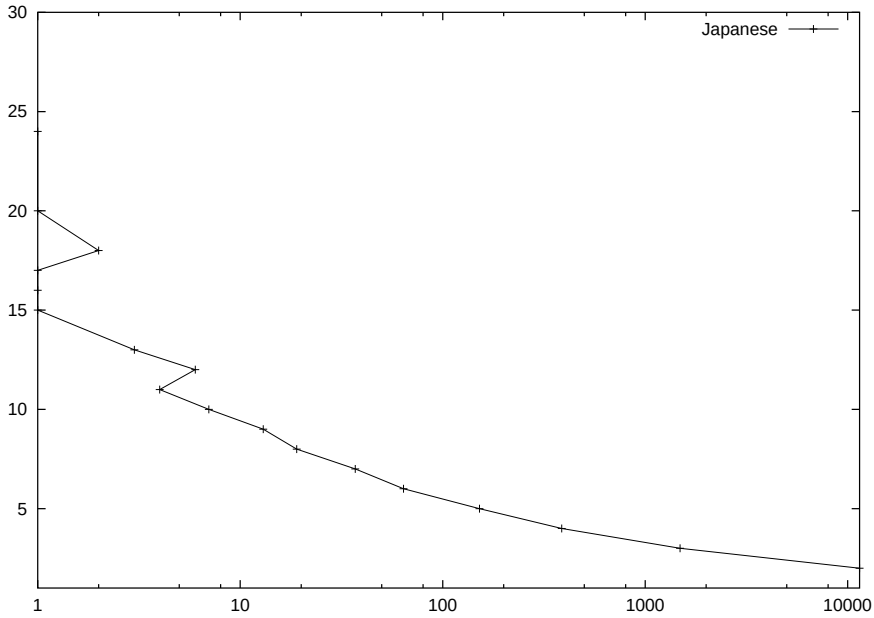

図 11: 日本語表現の抽出回数の分布 (1)

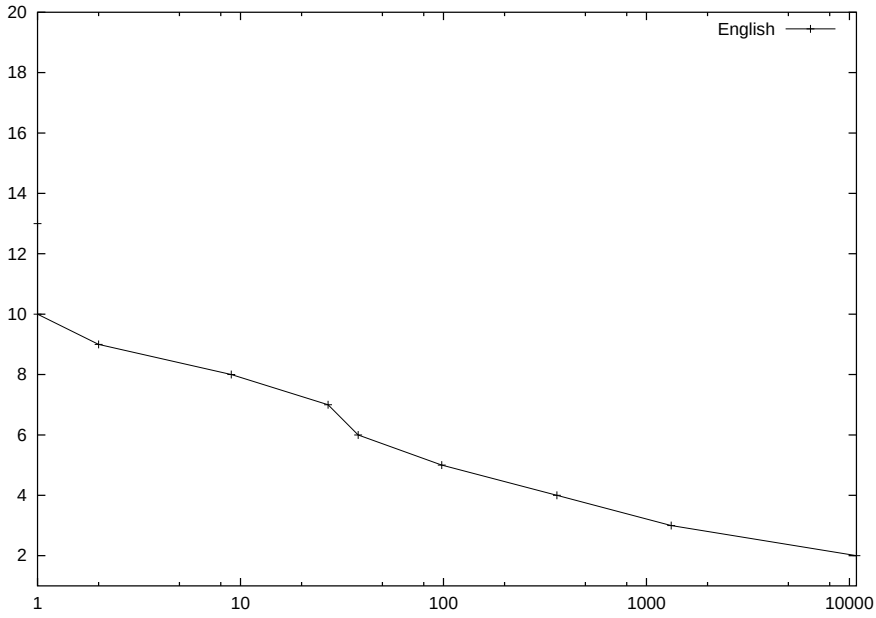

図 12: 英語表現の抽出回数の分布 (1)

抽出した日本語表現の一部を表 5 に，英語表現の一部を表 6 に示す．括弧内は表現の抽出回数で，抽出回数が多い順に 5 個掲載した．

表 5: 日本語表現の例 (1)

日本語表現
それは彼の (24)
彼女は私の (20)
あの人には (18)
彼の態度は (18)
彼らは彼を (17)

表 6: 英語表現の例 (1)

英語表現
I have a sore throat (13)
the lake is frozen over (10)
let's call it a day (9)
please make yourself at home (9)
I was at a loss for an answer (8)

日本語表現のうち抽出回数が一番多かった「それは彼の」を含む文と，英語表現のうち抽出回数が一番多かった「I have a sore throat」を含む文を対訳コーパスから調査した.「それは彼の」を含む文の一部を表 9 に,「I have a sore throat」を含む文の一部をを表 10 に示す．文番号が若い順に 5 個掲載した.

表 7: 日本文の例 (1)

文番号	日本文
879	それは彼の所有地の中にある.
895	それは彼の仕事にブレーキを掛けた.
1986	それは彼の思考を妨げた.
2156	それは彼の功績である.
3776	それは彼の感情を害した.

表 8: 英文の例 (1)

文番号	英語文
8499	I have a sore throat.
9506	I have a sore throat.
11642	I have a sore throat.
25304	I have a sore throat.
25502	I have a sore throat.

表 10 から，英語表現「I have a sore throat」自身が文として対訳コーパスに収録されていたことが分かった.「I have a sore throat」と対訳な日本文は,「私はのどが痛い.」,「のどが痛い.」のようなほぼ同じ意味を持つ文だった.

5.2.2 日英対訳パターンの候補の抽出

文番号を比較し，日本語表現と英語表現を対応付けた結果，2,735 個の日英対訳パターンの候補が抽出された (表 4)．日英対訳パターンの候補のうち，上位 5 個を表 9 に示す．ここで，“上位” とは，日英対訳パターンの候補を日本語表現の出現回数の多い順に並べた場合の順番である．以下，本稿では，日英対訳パターンの候補は日本語表現の出現回数の多い順に並んでいるものとし，“上位” は日英対訳パターンの候補のうち日本語表現の出現回数の多いもの，“下位” は日本語表現の出現回数の少ないものとする．

表 9: 日英対訳パターンの候補の例 (1)

日本語表現	英語表現	一致数	文番号の一致率
彼は動作が (8)	he is quick in action (5)	4	0.5(4/8)
お手紙ありがとう (6)	thank you for your letter (5)	4	0.7(4/6)
先んずれば人を制す (6)	he wins who gets the start (3)	3	0.5(3/6)
きょうはこれで (5)	let’s call it a day (9)	3	0.6(3/5)
悪事千里を走る (5)	ill news runs apace (4)	4	0.8(4/5)

ここで,「彼は動作が」と「he is quick in action」に対して，日本語表現を含む日本文および英語表現を含む英文を対訳コーパスから調査した．結果を表 10 に示す．

表 10: 対訳コーパスの調査の例 (1)

文番号	日本文	英文
124	彼は動作が すばやい.	he is quick in action.
3078	彼は動作が こそこそしている.	he is furtive in his movements.
21143	彼は動作が 敏しょうだ.	he is quick to act.
42938	彼は動作が 活発である.	he is quick in action.
43509	彼は動作が きびきびしている.	he is quick in action.
43513	彼は動作が 機敏だ.	he is quick in action.
47774	彼は動作が すばしっこい.	he is quickin his movements.
54621	彼は動作が 不活発だ.	he is slow in motion.
75769	動作が敏捷だ.	he is quick in action.

文番号は No.124, 42938, 43509, 43513 で一致していた．

5.3 80,000 文の名詞・動詞・副詞を置換した場合の結果

5.3.1 日本語表現および英語表現の抽出

実験では，まず対訳コーパスの日本文から日本語表現を 13,245 個抽出し，次に英文から英語表現を 19,985 個抽出した．日本語表現の抽出回数の分布を図 13 に，英語表現の抽出回数の分布を図 14 に示す．縦軸が表現の出現回数，横軸が表現の個数である．

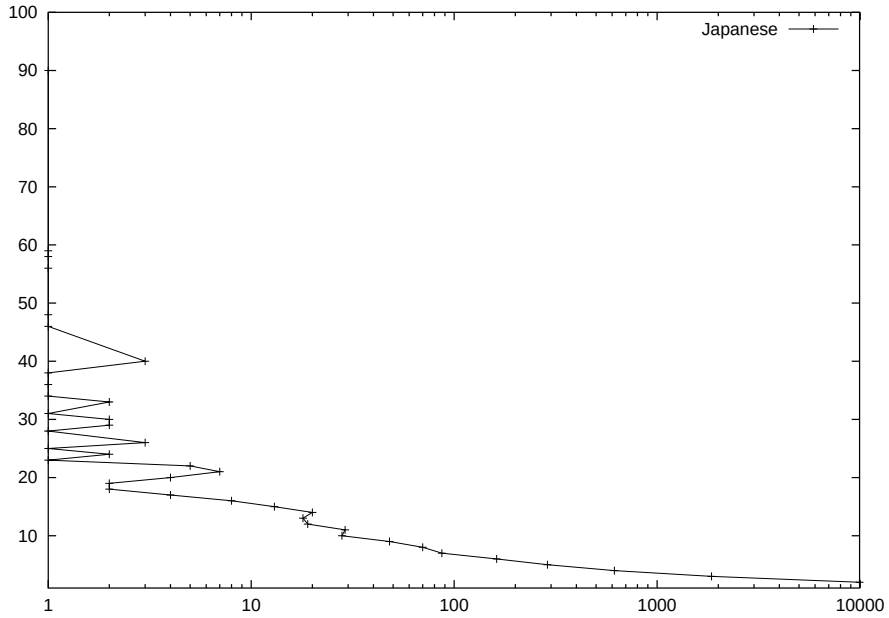

図 13: 日本語表現の抽出回数の分布 (2)

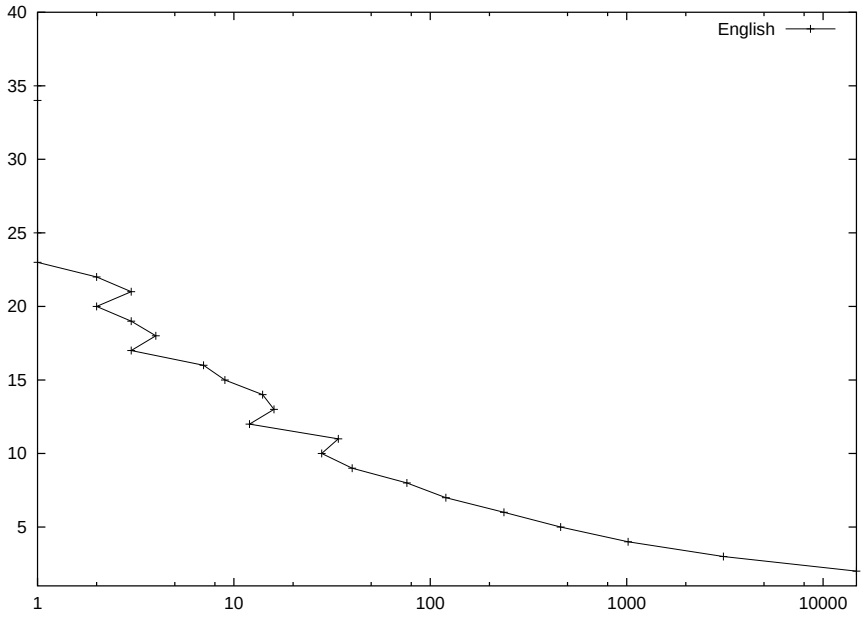

図 14: 英語表現の抽出回数の分布 (2)

抽出した日本語表現の一部を表 11 に，英語表現の一部を表 12 に示す．括弧内は表現の抽出回数で，抽出回数が多い順に 5 個掲載した．

表 11: 日本語表現の例 (2)

日本語表現
この N は N だ (90)
この N は N が V ている (59)
N は N を $AdvV$ た (58)
N は N を N で V た (56)
この N は N に V ている (48)

表 12: 英語表現の例 (2)

英語表現
the N is out of N (35)
the N is Adv in N (34)
the N V-ed in a N (25)
N V no N for N (23)
N V N on N's N (22)

日本語表現のうち抽出回数が一番多かった「この N は N だ」を含む文と，英語表現のうち抽出回数が一番多かった「the N is out of N」を含む文を対訳コーパスから調査した.「この N は N だ」を含む文の一部を表 13 に,「the N is out of N」を含む文の一部をを表 14 に示す．文番号が若い順に 5 個掲載した.

表 13: 日本文の例 (2)

文番号	日本文
11943	この N は N だ.
12767	この N は N だ.
12781	この N は N だ.
13464	この N は N だ.
15171	この N は N だ.

表 14: 英文の例 (2)

文番号	英語文
3412	the N is out of N.
3451	the N is out of N.
5815	the N is out of N.
5876	the N is out of N.
5877	the N is out of N.

表 13 と表 14 から，日本語表現「この N は N だ」自身と英語表現「the N is out of N」自身が文として対訳コーパスに収録されていたことが分かった.

5.3.2 日英対訳パターンの候補の抽出

文番号を比較し，日本語表現と英語表現を対応付けた結果，1,198 個の日英対訳パターンの候補が抽出された (表 4)．日英対訳パターンの候補のうち，上位 5 個を表 15 に示す．

表 15: 日英対訳パターンの候補の例 (2)

日本語表現	英語表現	一致数	文番号の一致率
この N は N で V ている (14)	this N is V-ed of N (18)	8	0.6(8/14)
N はこの N で N を V た (8)	this N V-ed N a N (7)	5	0.6(5/8)
この N は N だらけだ (8)	this N is full of N (15)	4	0.5(4/8)
N の N が Adv たい (7)	N is anxious to V the N of N's N (6)	5	0.7(5/7)
$AdvN$ が無い (6)	there is no N for more (5)	4	0.7(4/6)

ここで,「この N は N で V ている」と「this N is V-ed of N」に対して，日本語表現を含む日本文および英語表現を含む英文を対訳コーパスから調査した．結果の一部を表 16 に示す．文番号の若い順に 5 個掲載した．

表 16: 対訳コーパスの調査の例 (2)

文番号	日本文	英文
9526	この N は N で V ている.	this N is N Adv.
10774	この N は N で V ている.	this N is V-ed of N.
11570	この N は N で V ている.	this N is V-ed of N.
11943	この N は N だ.	this N is V-ed of N.
13689	この N は N で V ている.	this N is V-ed of N.

5.4 正解率

実験で抽出した日英対訳パターンの候補のうち，上位 50 個，ランダムに選んだ 50 個，下位 50 個に対して人手で評価を行い，正解率を求めた．それぞれの結果について述べていく．

5.4.1 上位 50 個の正解率

単文 10,000 文の結果を表 17 に，50,000 文の結果を表 18 に，80,000 文の結果を表 19 に示す．

表 17: 10,000 文の正解率 (上位 50 個)

	○	△	×	正解率	候補数
文字を単位とした字面	9	39	2	96%(48/50)	112
名詞の置換	6	42	2	96%(48/50)	133
名詞・動詞の置換	15	23	12	76%(38/50)	113
名詞・動詞・副詞の置換	11	26	13	74%(37/50)	108

表 18: 50,000 文の正解率 (上位 50 個)

	○	△	×	正解率	候補数
文字を単位とした字面	12	32	6	88%(44/50)	1,317
名詞の置換	14	28	8	84%(42/50)	1,058
名詞・動詞の置換	8	30	12	76%(38/50)	672
名詞・動詞・副詞の置換	9	26	15	70%(35/50)	613

表 19: 80,000 文の正解率 (上位 50 個)

	○	△	×	正解率	候補数
文字を単位とした字面	13	31	6	88%(44/50)	2,735
名詞の置換	12	33	5	90%(45/50)	2,181
名詞・動詞の置換	5	30	15	70%(35/50)	1,296
名詞・動詞・副詞の置換	4	31	15	70%(35/50)	1,198

全体では 10,000 文の文字を単位とした字面の場合と名詞を置換した場合で正解率が高かった．また，文数が増えてもほぼ同じ精度で日英対訳パターンが抽出できた．具体的には，10,000 文，50,000 文，80,000 文のそれぞれの場合で文字を単位とした字面の場合の正解率が高く，品詞の置換を行った場合の正解率が低くなる傾向にあった．

5.4.2 ランダムな 50 個の正解率

単文 10,000 文の結果を表 20 に，50,000 文の結果を表 21 に，80,000 文の結果を表 22 に示す.

表 20: 10,000 文の正解率 (ランダムな 50 個)

	○	△	×	正解率	候補数
文字を単位とした字面	10	37	3	94%(47/50)	112
名詞の置換	9	36	5	90%(45/50)	133
名詞・動詞の置換	13	29	8	84%(42/50)	113
名詞・動詞・副詞の置換	10	28	12	76%(38/50)	108

表 21: 50,000 文の正解率 (ランダムな 50 個)

	○	△	×	正解率	候補数
文字を単位とした字面	9	37	4	92%(46/50)	1,317
名詞の置換	7	35	8	84%(42/50)	1,058
名詞・動詞の置換	5	32	13	74%(37/50)	672
名詞・動詞・副詞の置換	5	29	16	68%(34/50)	613

表 22: 80,000 文の正解率 (ランダムな 50 個)

	○	△	×	正解率	候補数
文字を単位とした字面	14	34	2	96%(48/50)	2,735
名詞の置換	7	33	10	80%(40/50)	2,181
名詞・動詞の置換	6	36	8	84%(42/50)	1,296
名詞・動詞・副詞の置換	1	35	14	72%(36/50)	1,198

全体では 80,000 文の文字を単位とした字面の場合で正解率が高かった．また，前節と同様，文数が増えてもほぼ同じ精度で日英対訳パターンが抽出できた．具体的には，10,000 文，50,000 文，80,000 文のそれぞれの場合で文字を単位とした字面の場合の正解率が高く，品詞の置換を行った場合の正解率が低くなる傾向にあった.

5.4.3 下位 50 個の正解率

単文 10,000 文の結果を表 23 に，50,000 文の結果を表 24 に，80,000 文の結果を表 25 に示す.

表 23: 10,000 文の正解率 (下位 50 個)

	○	△	×	正解率	候補数
文字を単位とした字面	12	35	3	94%(47/50)	112
名詞の置換	11	34	5	90%(45/50)	133
名詞・動詞の置換	5	39	6	88%(44/50)	113
名詞・動詞・副詞の置換	4	35	11	78%(39/50)	108

表 24: 50,000 文の正解率 (下位 50 個)

	○	△	×	正解率	候補数
文字を単位とした字面	10	38	2	96%(48/50)	1,317
名詞の置換	2	39	9	82%(41/50)	1,058
名詞・動詞の置換	3	36	11	78%(39/50)	672
名詞・動詞・副詞の置換	4	40	6	88%(44/50)	613

表 25: 80,000 文の正解率 (下位 50 個)

	○	△	×	正解率	候補数
文字を単位とした字面	22	27	1	98%(49/50)	2,735
名詞の置換	2	41	7	86%(43/50)	2,181
名詞・動詞の置換	3	38	9	82%(41/50)	1,296
名詞・動詞・副詞の置換	3	39	8	84%(42/50)	1,198

全体では 80,000 文の文字を単位とした字面の場合で正解率が高かった．また，前節，前々節と同様，文数が増えてもほぼ同じ精度で日英対訳パターンが抽出できた．具体的には，10,000 文，50,000 文，80,000 文のそれぞれの場合で文字を単位とした字面の場合の正解率が高く，品詞の置換を行った場合の正解率が低くなる傾向にあった.

5.5 日英対訳パターンの例

完全に対訳であると判断した評価 “◯” の例を表 26 に，ほぼ対訳であると判断した評価 “△” の例を表 27 に，対訳ではないと判断した評価 “×” の例を表 28 に示す (詳しくは付録参照)．表 26 や表 27 から，意味的にまとまりを持つ日英対訳パターンを抽出できたことが分かった．

表 26: 評価 “◯” の例

	日本語表現	英語表現	一致数	一致率
文字を単位とした字面	お手紙ありがとう (6)	thank you for your letter (5)	4	0.7(4/6)
	悪事千里を走る (5)	ill news runs apace (4)	4	0.8(4/5)
名詞の置換	この N は N で できている (9)	this N is made of N (16)	8	0.9(8/9)
	この N は N だらけだ (8)	this N is full of N (15)	4	0.5(4/8)
名詞・動詞の置換	この N は N で V ている (14)	this N is V-ed of N (18)	8	0.6(8/14)
	N が N 無く V だ (4)	the N V-ed without a N (4)	2	0.5(2/4)
名詞・動詞・副詞の置換	N の N は Adv V れている (3)	N's N is V-ed Adv (3)	2	0.7(2/3)
	この N には Adv N の N が V (4)	a N of the N Adv V about this N (2)	2	0.5(2/4)

表 27: 評価 “△” の例

	日本語表現	英語表現	一致数	一致率
文字を単位とした字面	彼は動作が (8)	he is quick in action (5)	4	0.5(4/8)
	詩を朗読した (5)	he recited a poem (3)	3	0.6(3/5)
名詞の置換	N に N が見えた (7)	N saw a N in the N (4)	4	0.6(4/7)
	N は N を破った (6)	N broke N's N (8)	3	0.5(3/6)
名詞・動詞の置換	正しい N を V た (2)	N V-ed the right N (4)	2	1(2/2)
	膨大な N に V (2)	N V to a enormous N (3)	2	1(2/2)
名詞・動詞・副詞の置換	$AdvN$ の N が悪い (4)	N is wrong with N's N (4)	2	0.5(2/4)
	無私の N は Adv 尊い (2)	N is Adv precious (2)	2	1(2/2)

表 28: 評価 “×” の例

	日本語表現	英語表現	一致数	一致率
文字を単位とした字面	あの方には (4)	deeply indebted to him (3)	2	0.5(2/4)
	協議には、 (4)	the consultations should include a discussion on (2)	2	0.5(2/4)
名詞の置換	N が N 斤 (4)	N have put on (4)	2	0.5(2/4)
	但し N の N だ (2)	N is a questionable N (5)	2	1(2/2)
名詞・動詞の置換	N 天を V (6)	the V-ing N V sky-high (3)	3	0.5(3/6)
	N 名を V (5)	a N V a N behind N (4)	3	0.6(3/5)
名詞・動詞・副詞の置換	N が $AdvV$ ぬ (6)	N can Adv V the N (3)	3	0.5(3/6)
	N が Adv ぬ (6)	N can Adv V the N V (3)	3	0.5(3/6)

ここで，表 26 の文字を単位とした字面の場合で抽出した「お手紙ありがとう」と「thank you for your letter」に対して，日本語表現を含む日本文および英語表現を含む英文を対訳コーパスから調査した．結果を表 29 に示す．他の日英対訳パターンの例は付録参照.

表 29: 日本語表現および英語表現を含む文の例 (1)

文番号	日本文	英文
2933	お手紙ありがとう.	thank you for your letter.
9373	お手紙ありがとう.	thank you for your letter.
10934	お手紙ありがとう.	thank you for your letter.
18815	お手紙ありがとう.	thank you for your letter.
20279	お手紙拝見しました .	thank you for your letter.
50730	ご丁寧な お手紙ありがとう.	thank you for your polite letter.
50759	お手紙ありがとう.	thank you for your kind letter.

文番号は No.2933, 9373, 10934, 18815 で一致していた.

6　考察

6.1　連鎖型共起表現を対象にした日英対訳パターン

6.1.1　正解率の推移

品詞の置換を行った日英対訳パターンは，字面の場合に比べ正解率が低かった．品詞の置換を行うことで字面の情報が失われ，日本語表現と英語表現の意味的対応の判断が困難だったためだと考えられる．しかし，様々な種類の日英対訳パターンが必要となる場合があるため，品詞の置換を行った日英対訳パターンは，今後，日英対訳パターンを作成する上で有効な指針になることが期待できる．

6.1.2　日英対訳パターンの作成

本実験において，ほぼ対訳であると判断した評価 “△” の日英対訳パターンが多かった．評価 “△” の日英対訳パターンは，人手で修正を行うことで完全に対訳な日英対訳パターンを作成することができる．

評価 “△” の日英対訳パターンを人手で修正し日英対訳パターンを作成する場合，人手での修正は困難な場合があるが，日英対訳パターンの作成を補助できると考えられる．

表 27 に対して人手で修正を行った結果を表 30 に示す．下線部分が修正箇所である．

表 30: 評価 “△” の修正の例

	日本語表現	英語表現
文字を単位とした字面	彼は動作が すばやい	he is quick in action
	彼は 詩を朗読した	he recited a poem
名詞の置換	N はN に N が見えた	N saw a N in the N
	N は N のN を破った	N broke N’s N
名詞・動詞の置換	N は 正しい N を V た	N V-ed the right N
	N は 膨大な N に V	N V to a enormous N
名詞・動詞・副詞の置換	N は$Adv N$ の N が悪い	N is Adv wrong with N’s N
	N は Adv 尊い (“無私の” を削除)	N is Adv precious

6.2 離散型共起表現を対象にした日英対訳パターン

6.2.1 日英対訳パターンの例

本手法では,「betweeen 〜 and」のような離れた場所にある離散型共起表現を対象にした日英対訳パターンの抽出も可能である.

単文 80,000 文の文字を単位とした字面の場合の弱抑制型では，日英対訳パターンの候補を 3 個抽出した．結果を表 31 に示す．また，強抑制型では，日英対訳パターンの候補を 3 個抽出した．結果を表 32 に示す．弱抑制型と強抑制型では，同じ日英対訳パターンを抽出した.

表 31: 離散型共起表現の結果 (弱抑制型)

評価	日本語表現	英語表現	一致数	一致率
△	は、いくつかの点で 〜と異なる (2)	differs from 〜 in several ways (2)	2	1(2/2)
△	大佐は〜位が下である (2)	a colonel is 〜 a general (2)	2	1(2/2)
○	翻訳は原文の 〜を存している (2)	the translation retains the 〜 of the original (4)	2	1(2/2)

表 32: 離散型共起表現の結果 (強抑制型)

評価	日本語表現	英語表現	一致数	一致率
△	は、いくつかの点で 〜と異なる (2)	differs from 〜 in several ways (2)	2	1(2/2)
△	大佐は〜位が下である (2)	a colonel is 〜 a general (2)	2	1(2/2)
○	翻訳は原文の 〜を存している (2)	the translation retains the 〜 of the original (4)	2	1(2/2)

離散型共起表現の弱抑制型で抽出した評価 “○” の「翻訳は原文の～を存している」と「the translation retains the ～ of the original」に対して，日本語表現を含む日本文および英語表現を含む英文を対訳コーパスから調査した．結果を表 33 に示す．

表 33: 日本語表現および英語表現を含む文の例 (2)

文番号	日本文	英文
62741	訳文は原文の趣きを失わぬ.	the translation retains the flavour of the original.
65931	翻訳は原文の 口調 を存している.	the translation retains the one of the original.
66723	訳文は原文の妙味を存している.	the translation retains the charm of the original.
74330	翻訳は原文の 調子 を存している.	the translation retains the tone of the original.

文番号は No.65931, 74330 で一致していた．

6.2.2 日本語表現と英語表現の数

離散型共起表現を対象にした日英対訳パターンは，連鎖型共起表現を対象にした場合に比べ，抽出した日英対訳パターンの候補数が少なかった．そこで，単文 80,000 文の字面の場合に対して，連鎖型共起表現を対象にした場合と離散型共起表現を対象にした場合の比較を行った．結果を表 34 に示す．

表 34: 連鎖型共起表現と離散型共起表現の比較

	日本語表現の数	英語表現の数	候補数
連鎖型共起表現	13,629	12,691	2,735
離散型共起表現 (弱抑制型)	1,331	318	3
離散型共起表現 (強抑制型)	1,016	252	3

表 34 から，離散型共起表現を対象とした場合は連鎖型共起表現を対象とした場合と比較すると，対訳コーパスから別々に抽出した日本語表現と英語表現の数が少ないことが分かった．また，離散型共起表現を対象とした場合は，強抑制型よりも弱抑制型の方が日本語表現と英語表現の数が多かった．今後は，離散型共起表現を増やすため，日本語表現および英語表現をさらに効率良く抽出する必要があると考えられる．

6.2.3 離散型共起表現の弱抑制型と強抑制型

離散型共起表現の弱抑制型と強抑制型では同じ日英対訳パターンを抽出した (6.2.1 節). 同じ日英対訳パターンを抽出したことについての考察を行う．対訳コーパスから別々に抽出した離散型共起表現の日本語表現と英語表現の例を示す．弱抑制型の例を表 35 と表 36 に，強抑制型の例を表 37 と表 38 に示す.

表 35: 離散・弱抑制型の例 (日本語表現)

日本語表現
これは〜です (9)
には〜がある (5)
彼らは〜に出かけた (5)
にはまだ〜が残っている (4)
でも〜何でも (2)

表 36: 離散・弱抑制型の例 (英語表現)

英語表現
spacing on 〜 axis (3)
they are on 〜 terms with each other (3)
I beg you will 〜 favour (2)
I checked my bag 〜 the cloakroom (2)
I want to change 〜 division (2)

表 37: 離散・強抑制型の例 (日本語表現)

日本語表現
これは〜です (9)
には〜がある (5)
彼らは〜に出かけた (5)
にはまだ〜が残っている (4)
これは〜だ (4)

表 38: 離散・強抑制型の例 (英語表現)

英語表現
they are on 〜 terms with each other (3)
spacing on 〜 axis (2)
I beg you will 〜 favour (2)
I checked my bag 〜 the cloakroom (2)
I concentrated 〜 my energies on (2)

表 35 の日本語表現の「でも〜何でも」と，表 36 の英語表現の「I want to change 〜 division」は弱抑制型のみから抽出された．残りの表現は弱抑制型と強抑制型のどちらからも抽出された．本実験では，離散型共起表現を対象にした場合の弱抑制型と強抑制型では同じ日英対訳パターンを抽出した (6.2.1 節)．しかし，弱抑制型のみから抽出される日本語表現および英語表現が存在するため，6.2.1 節で弱抑制型と強抑制型の日英対訳パターンが同じだったことは偶然の一致で，弱抑制型のみから抽出される日英対訳パターンが存在する可能性があると考えられる．

7　おわりに

本研究では，人手での日英対訳パターンの作成を補助するため，まず，N-gram 統計処理方法を用いて，対訳コーパスから日本語表現および英語表現を抽出した．次に，文番号の一致率を求めることで日英対訳パターンの候補を自動抽出した．そして，抽出した日英対訳パターンの候補を人手で評価し，正解率を求めた．

連鎖型共起表現を対象に，品詞の置換を行い実験を行った結果，様々な種類の意味的にまとまりを持つ日英対訳パターンの候補を自動抽出することができた．それぞれの日英対訳パターンの候補数と正解率を求めた結果，文字を単位とした字面の場合が抽出した日英対訳パターンの候補が多く，正解率も高かった．また，全ての正解率の平均を求めたところ，約 84%の候補が人手でによる判断で日英対訳パターンを作成できることが分かった．抽出した日英対訳パターンは，ほぼ対訳であると判断した評価 “△” が多かった．評価 “△” の日英対訳パターンは人手で修正を行うため，今後，評価 “△” が減り，完全に対訳であると判断される日英対訳パターンが増えれば，効率良く日英対訳パターンを作成することができると考えられる．

一方，離散型共起表現を対象に実験を行った結果，自動抽出した日英対訳パターンの候補は少なかった．今後は，離散型共起表現を対象にした日英対訳パターンの候補を増やす必要があると考えられる．

謝辞

最後に，本研究において御指導頂きました鳥取大学工学部知能情報工学科計算機 C 研究室の池原教授，村上助教授，徳久助手に厚くお礼申し上げます.

また，本研究に使用させて頂いた論文，本の著者の方々にお礼申し上げます.

参考文献

[1] 熊野明, 平川秀樹, “対訳文書からの機械翻訳専門用語辞書作成”, 情報処理学会論文誌, Vol.35, No.11, pp.2283-2290, 1994.

[2] 北村美穂子, 松本裕治, “対訳コーパスを利用した対訳表現の自動抽出”, 情報処理学会論文誌, Vol.38, No.4, pp727-736, 1997.

[3] 池原悟, 白井諭, 河岡司, “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”, 情報処理学会論文誌, Vol.36, No.11, pp.2584-2596, 1995.

[4] 波宏誠, “N-gram 統計を応用した日本語共起表現辞書の作成”, 鳥取大学工学部知能情報工学科卒業論文, 1999.

[5] 道祖尾太祐, 村上仁一, 徳久雅人, 池原悟, “日英対訳パターンの自動抽出に向けて”, 情報処理学会研究報告, 2003-NL-153, pp.113-118, 2003.

[6] 村上仁一, “英日対訳データベースの現状”, 「言語, 認識, 表現」第 7 回年次研究会プログラム, 2002.

[7] 斎藤健太郎, “大規模コーパスからの重文複文の統語構造の自動抽出”, 鳥取大学工学部知能情報工学科卒業論文, 2000.

[8] 道祖尾太祐, 村上仁一, 徳久雅人, 池原悟, “N-gram を利用した日英対訳パターンの自動抽出”, 言語処理学会第 10 回年次大会発表予定, 2004-3.