自然言語処理　138－7
（2000．7．18）

対訳コーパスにおける低頻度語の性質
訳語対自動抽出に向けた基礎研究

辻　慶太[1)]　芳鐘冬樹[2)]　影浦　峡[3)]

東京大学大学院教育学研究科[1)2)]
〒113-0033 東京都文京区本郷 7-3-1
i34188@m-unix.cc.u-tokyo.ac.jp[1)]
fuyuki@p.u-tokyo.ac.jp[2)]

国立情報学研究所[3)]
〒101-8430 東京都千代田区一ツ橋 2-1-2
kyo@rd.nacsis.ac.jp[3)]

あらまし　既に辞書に載っている訳語対を，対訳コーパスから自動抽出してもメリットは少ない。コーパス中の頻度が高い対は既に辞書に載っているであろう。対訳コーパスから自動抽出すべき訳語対は，頻度の低い訳語対である。そのような前提から本研究では，これまで研究されてきた統計的な訳語対抽出手法では，低頻度訳語対の抽出が難しいことを示す。具体的には，統計的手法では同じ言語の２語が常に同じアラインメントに共起する場合，訳語が決定できない問題を取り上げる。頻度の低い語同士はこうした決定不能状況に陥りやすい。本研究では，実際の対訳コーパス中で決定不能状況にある低頻度語の量・質を調べ，訳語対抽出手法の改善方向を検討した。

キーワード　低頻度語，訳語対，自動抽出，対訳コーパス

Low-frequency Words in Bilingual Corpora
A Step towards Automatic Extraction of Bilingual Word Pairs

Keita Tsuji[1)]　Fuyuki Yoshikane[2)]　Kyo Kageura[3)]

Graduate School of Education,
University of Tokyo[1)2)]
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033
i34188@m-unix.cc.u-tokyo.ac.jp[1)]
fuyuki@p.u-tokyo.ac.jp[2)]

National Institute of Informatics[3)]
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
kyo@rd.nacsis.ac.jp[3)]

Abstract　The high-frequency bilingual word pairs in bilingual corpora are already listed in the dictionaries. It is the low-frequency pairs that we have to extract. Based on that idea, we examine the method for automatically extracting bilingual word pairs from corpora and show that the statistical method, which has been studied intensively so far, is not suitable for the task. If two words *J1* and *J2* which belong to the same language always co-occur in the same alignments, the statistical method cannot determine which word is the correct translation of word *E* which belong to the other language. We saw many of the low-frequency words are in the above situation.

key words　Low-frequency word, Bilingual word pair, Automatic extraction, Bilingual Corpora

1 はじめに

近年，対訳コーパスの入手可能性が高まるにつれ，そこから訳語対を自動抽出する研究が盛んに行われている。一般に，対訳コーパスが存在する状況・分野には，英和辞書や学術用語集のような，二言語辞書・対訳用語リストが存在すると考えるのが自然である。そうした辞書・用語リストに載っている訳語対を，対訳コーパスからあらためて抽出してもメリットは少ない。対訳コーパスから自動抽出する価値のある訳語対は，少なくとも辞書・用語リストに載っていない訳語対である。

コーパス中の頻度が高い訳語対は，一般的あるいは分野の中心的な訳語対であり，既に辞書や用語リストに載っている可能性が高い。逆に頻度が低い訳語対は載っていない可能性が高い。以上のような見地から，対訳コーパスからの訳語対自動抽出で重要なのは，低頻度訳語対の抽出であると考える[1]。本研究は，低頻度訳語対は先行研究手法の多くでは抽出が難しいことを示し，改善の方向を検討するものである。

対訳コーパスからの訳語対自動抽出手法は，利用する情報の観点から次の３つに分類できる。即ち，(1) 訳文や対訳抄録といった，意味的対応のある単位（以下「アラインメント」と呼ぶ）をまずコーパスに設定し，語がアラインメントに出現・共起する頻度を利用する手法 [1][2][3][4][5][6][7][8][9][10][11][12][13][14]，(2) 辞書が挙げる訳語対あるいはそれらを組み合わせた形との表記上の類似を利用する手法 [4][9][15][16][17]，(3) 借用語と元の語に見られるような，音形上の類似を利用する手法 [17][18][19]，の３つである。

辞書収録語に関する先述の傾向から，(2) は少なくとも単位語に関する低頻度訳語対の抽出には向いていない。(1) の統計的手法に関しては，次の問題を指摘できる。即ち，同じ言語の２語 J_1 と J_2 があり，J_1 に対する訳語が E であったとする。統計的手法では，J_1 と J_2 が，常に同じアラインメントに現れる場合，いずれが E の訳語なのか決定できない。例えば「変調」と「補数」の出現頻度が１で，かつ出現アラインメントが同じであった場合，いずれが 'modulation' の訳語なのか，統計的手法では決定できない。低頻度の語同士は，このような状況（以下「決定不能状況」と呼ぶ）に陥りやすい。

以下では，低頻度語の代表として頻度１の語を取り上げ，それらがどの程度コーパス中で決定不能状況にあり，(1) の手法では抽出が難しくなっているかを，実際のデータに基づいて示す。さらに状況改善の方法を考え，低頻度訳語対抽出に向けて (3) も含めたいくつかの方向を提案する。

[1] 例えば新語の訳語対は，本研究が実験に用いた対訳コーパスのように，更新の早いコーパスには含まれているが，辞書や用語リストには一般に含まれていない。新語は，コーパスにおける頻度が本来的に低い。

2 実験

本研究では，アラインメントの大きさ及びサンプルサイズが異なる４つのコーパスを取り上げ，それぞれにおける頻度１語を調査対象とした。頻度１語としては，次の３種類を調べた。即ち，(1) 名詞１単位，(2) 名詞の２単位連続，(3) 名詞１単位の専門用語，である。

2.1 調査対象としたコーパス

国立情報学研究所の学会発表データベースから，まず以下のデータを切り出し，調査対象対訳コーパスとした。即ち，(a) 人工知能分野の日英抄録 1,000 対，(b) 情報処理分野の日英抄録 9,084 対，(c) 情報処理分野の日英タイトル 25,533 対，である（以下それぞれ「AI 抄録」「IP 抄録」「IP タイトル」と略す）。これら３コーパスでは，抄録・タイトルをアラインメントとした。

さて，統計的な訳語対抽出手法では，文単位に対応した対訳コーパスを用いて，文をアラインメントにする場合が多い [3][6][7][8][9][10][11][12][13][14]。アラインメントの取り方は，頻度に大きく影響する。文をアラインメントにした場合も調べたい。ところで本研究の実験では，アラインメントの他方の言語の側に訳語が存在することを仮定して，一方の言語の側を調べるだけで，コーパスの対訳性は実際には（一部の実験を除いて）利用しない。従って，単一言語コーパスに対して，文をアラインメント（語の頻度を数える単位）として調査を行っても，対訳コーパス (a)(b)(c) における調査と，結果の比較はある程度可能と考えた。いわば単一言語コーパスを，仮想的な対訳コーパスの一言語側とみなしたと言える。そのようなコーパスとして，本研究では (b) を日本語部分と英語部分に分けたも

のを用い（以下「IP 文」と略す），それらにおいては文を単位として語の頻度を数えた。日本語は42,058 文，英語は 46,910 文から成る。

4 コーパスはそれぞれ，日本語側は茶筌 2.0 で，英語側は Brill tagger で形態素解析した。以下では，(1) 名詞 1 単位，(2) 名詞の 2 単位連続，(3) 名詞 1 単位の専門用語，の頻度 1 語を取り上げ，決定不能状況を調べたい。

2.2 名詞 1 単位

ここで名詞とは，日本語は茶筌が「名詞-一般」「名詞-サ変接続」と判定した単語（「システム」「推論」など），英語は Brill tagger が NN，NNP，NNS と判定した単語（'systems'，'reasoning' など）とした。これら名詞 1 単位の延べ・異なり数は，表 1 の通りである。次にこれらを，頻度ごとに分類すると表 2 のようになった。括弧内の数値は，全異なり語に占める割合（%）である。例えば，IP 抄録の日本語部分には，頻度 1 の語が 8,058 個あり，これらが全異なり語に占める割合は 49.27 % であることが分かる。4 コーパスすべてにおいて，全異なり語の約半分が頻度 1 となっている一方，頻度 5 以上の語は全体の 2，3 割にとどまっている。ここで取り上げた 4 コーパス，少なくとも IP 抄録・IP 文は，訳語対抽出源として入手可能な現実の対訳コーパスに比べ，特に小さい訳ではない。一般に，コーパスには低頻度語が多い。

次に頻度 1 語を何個含むかという観点で，アラインメントを分類した。結果は表 3 のようになった。表 3 から例えば，IP 抄録の日本語部分において頻度 1 の異なった語を 3 個含むアラインメントは 456 個あり，全アラインメント 9,084 個に占める割合は 5.02 % であることが分かる。先述のように，頻度 1 の語が同一アラインメントに複数現れている場合，統計的手法では訳語を決定することができない。そのような語がいくつあるかは，この表から算出できる。例えば IP 抄録の日本語部分においては，$1,045 \times 2 + 456 \times 3 + 225 \times 4 + 101 \times 5 ... = 5,658$ 個の頻度 1 語が，決定不能になっている。4 コーパスそれぞれに，こうした決定不能語がいくつあるかを調べたところ，表 4 のようになった。

表 4 から，例えば AI 抄録日本語部分においては，2,031 個の頻度 1 語が決定不能になっており，全異なり語に占める割合は 45.93 % にのぼることが分かる。一方，IP 文では，日本語英語とも決定不能語は全異なり語の 16 % 程度にとどまっている。文対応の対訳コーパスを用いれば，決定不能状況は深刻でないとも言えるが，これについては次の 2 点を検討する必要がある。即ち，(1) 抄録程度の単位で対応した対訳コーパスは比較的多いが，文単位で対応した対訳コーパスは少なく，現実的な訳語対抽出源として入手可能性が低い，(2) ここでは全異なり語に占める割合を示したが，冒頭で述べたように，抽出対象としては低頻度語の方が重要である。仮に頻度 1 語に占める決定不能語の割合を考えると，IP 文においてもその割合は 3 割を超える，の 2 点である。

頻度 1 の語は，このように決定不能状況に陥りやすい。その改善に向け，次の 3 処理の有効性を調べてみた。

(A) 辞書に載っている訳語対は，抽出する必要のない訳語対であるとして，アラインメントから除外する。これにより頻度 1 語を減らせるかもしれない。冒頭で，低頻度語は辞書に載っていないと仮定したが，ここでその検証を兼ねたい。

(B) 日本語のカタカナ語を正規化し，英語の複数形は単数形に統一する。これにより「インターフェース」と「インタフェース」，'systems' と 'system' のような同語異形が統一され，頻度 1 語を減らせるかもしれない。

(C) カタカナ語は，翻字規則により英訳語をある程度決定できる。この前提のもとに，カタカナ語やローマ字語は日本語部分から除外する。

2.2.1 辞書訳語対の除外

辞書に訳語対として載っている語同士を，アラインメントから除外した上で，再び決定不能状況を調べた。実験では，一般辞書と専門用語辞書の 2 種類を用意し，一方のみを用いた場合，両者を併用した場合を調べた。一般辞書には EDICT を，専門用語辞書には，人工知能分野に関しては『人工知能大辞典』（以下 'AD'），情報処理分野に関しては『英和コンピュータ用語大辞典第 2 版』（以下 'CD'）を

		AI 抄録	IP 抄録	IP タイトル	IP 文
	アラインメント数	1,000	9,084	25,533	日本語 42,058 文 英語 46,910 文
日本語	延べ語数	55,779	467,270	159,392	467,270
	異なり語数	4,422	16,354	10,417	16,354
英語	延べ語数	38,533	302,936	133,586	302,936
	異なり語数	4,151	16,629	14,050	16,629

表 1: 名詞 1 単位に関する延べ・異なり数

		AI 抄録		IP 抄録		IP タイトル		IP 文	
日本語	全異なり語数	4,422		16,354		10,417		16,354	
	頻度 1 の異なり語	2,269	(51.31)	8,058	(49.27)	5,080	(48.77)	6,697	(40.95)
	頻度 2 〃	644	(14.56)	2,236	(13.67)	1,516	(14.55)	2,446	(14.96)
	頻度 3 〃	278	(6.29)	1,123	(6.87)	737	(7.07)	1,410	(8.62)
	頻度 4 〃	190	(4.30)	737	(4.51)	476	(4.57)	829	(5.07)
	頻度 5 以上 〃	1,041	(23.54)	4,200	(25.68)	2,608	(25.04)	4,972	(30.40)
英語	全異なり語数	4,151		16,629		14,050		16,629	
	頻度 1 の異なり語	2,206	(53.14)	9,339	(56.16)	8,191	(58.30)	7,527	(45.26)
	頻度 2 〃	590	(14.21)	2,113	(12.71)	1,875	(13.35)	2,564	(15.42)
	頻度 3 〃	295	(7.11)	999	(6.01)	858	(6.11)	1,326	(7.97)
	頻度 4 〃	174	(4.19)	610	(3.67)	531	(3.78)	839	(5.05)
	頻度 5 以上 〃	886	(21.34)	3,568	(21.46)	2,595	(18.47)	4,373	(26.30)

表 2: 名詞 1 単位に関する頻度毎の異なり数

用いた[2]。結果は表 5 のようになった[3]。表 5 から例えば，IP 抄録の日本語部分では，EDICT に含まれる訳語対を除外することで，決定不能語が 5,658 個から 5,334 個に減り，全異なり語に占める割合は 32.62 %になることが分かる。

全体として，辞書に含まれる訳語対を除いても，頻度 1 の決定不能語はあまり減らない。これは逆に，低頻度語には辞書にない語が多いこと，即ち，冒頭で述べた仮説が一定の妥当性を持つことをも示している。

2.2.2 カタカナ語の正規化と英語の単数形化

次に，4 文字以上のカタカナ語では長音記号「ー」をなくす，エ段の音 E +「イ」は E にする，といったカタカナ語の正規化と，英語複数形の単数形化に関する有効性を検証した。これらは，必要に応じ元の語形が再生できる範囲内で，記号化・統一を図ったと言える。結果は，表 6 のようになった。全般に，カタカナ語処理・複数形処理は若干の決定不能語減少をもたらすものの，決定的な力は見られない。

2.2.3 カタカナ語・ローマ字語の除外

頻度 1 語の語種を調べたところ，表 7 のようになり，頻度 1 語の半分以上は，カタカナ語・ローマ字語であることが分かった。カタカナ語は，翻字規則によって訳語を決定できる可能性があり [17][19]，ローマ字語についてもそれは言える[4]。つまり，カタカナ語・ローマ字語は，統計的抽出手法でなく他の手法に抽出を託す，という考えが成り立つ。そのような前提のもとで，カタカナ語・ローマ字語を除去し，決定不能語削減を試みた。結果は表 8 の

[2] それぞれの異なり訳語対数は，EDICT(102,380 対)，AD(3,738 対)，CD(38,549 対)，AD + EDICT(105,807 対)，CD + EDICT(138,838 対) であった。

[3] 本実験ではコーパスの対訳性を利用するので，IP 文は調査対象としていない。

[4] そもそも日本語のローマ字語と英語は抽出に値する訳語対なのかという疑問もある。

		AI 抄録		IP 抄録		IP タイトル		IP 文	
日本語	0	237	(23.70)	4,751	(52.30)	21,288	(83.37)	36,905	(87.75)
	1	238	(23.80)	2,400	(26.42)	3,575	(14.00)	4,052	(9.63)
	2	180	(18.00)	1,045	(11.50)	541	(2.12)	804	(1.91)
	3	118	(11.80)	456	(5.02)	103	(0.40)	205	(0.49)
	4	81	(8.10)	225	(2.48)	18	(0.07)	63	(0.15)
	5	54	(5.40)	101	(1.11)	6	(0.02)	18	(0.04)
	6	31	(3.10)	47	(0.52)	2	(0.01)	4	(0.01)
	7	21	(2.10)	27	(0.30)	0	(0.00)	3	(0.01)
	8	18	(1.80)	12	(0.13)	0	(0.00)	2	(0.00)
	9	8	(0.80)	4	(0.04)	0	(0.00)	1	(0.00)
	10+	14	(1.40)	16	(0.18)	0	(0.00)	1	(0.00)
	計	1,000	(100.00)	9,084	(100.00)	25,533	(100.00)	42,058	(100.00)
英語	0	230	(23.00)	4,301	(47.35)	18,859	(73.86)	40,936	(87.26)
	1	245	(24.50)	2,526	(27.81)	5,424	(21.24)	4,876	(10.39)
	2	197	(19.70)	1,205	(13.27)	1,039	(4.07)	800	(1.71)
	3	116	(11.60)	515	(5.67)	165	(0.65)	203	(0.43)
	4	85	(8.50)	251	(2.76)	38	(0.15)	61	(0.13)
	5	36	(3.60)	125	(1.38)	7	(0.03)	20	(0.04)
	6	32	(3.20)	66	(0.73)	0	(0.00)	4	(0.01)
	7	22	(2.20)	43	(0.47)	1	(0.00)	6	(0.01)
	8	13	(1.30)	17	(0.19)	0	(0.00)	4	(0.01)
	9	12	(1.20)	10	(0.11)	0	(0.00)	0	(0.00)
	10+	12	(1.20)	25	(0.28)	0	(0.00)	0	(0.00)
	計	1,000	(100.00)	9,084	(100.00)	25,533	(100.00)	46,910	(100.00)

表 3: 頻度 1 の名詞 1 単位を含む個数ごとのアラインメント数

ようになった。表 8 から例えば，IP 抄録日本語側において，5,658 個あった決定不能語が，カタカナ語・ローマ字語の除外で 1,112 個になり，全異なり語の 6.80 %にまで減ったことが分かる。決定不能語削減に，カタカナ語・ローマ字語の除外は有効である。

2.3 名詞の 2 単位連続

前節までは，名詞 1 単位を扱ってきた。次に名詞の 2 単位連続（「情報検索」「並列処理」など）を対象に，これまで同様の調査を行った。結果は表 9, 10 のようになった。表 9 から，名詞の 2 単位連続においては，決定不能語が非常に多くなること，比較的決定不能語の少なかった IP タイトルや IP 文においても，その量は無視しがたいものになること，表 10 から，カタカナ・ローマ字語の除外も大きな効果を上げなくなること，が分かる。対訳コーパスから，統計的手法に基づいて複合語の訳語対を抽出する場合は，複合語を 1 つの固まりとして扱うのは避け，語構成要素同士の対応といった，より小さい単位での対訳性を考えるべきであろう。

2.4 名詞 1 単位の専門用語

先行研究の中には，訳語対のうち専門用語の訳語対を，主要な抽出対象とするものがある。アラインメント中の専門用語を何らかの方法で特定し，候補語を専門用語に限定すれば，決定不能語は減少するかもしれない。例えば「補数」と「畢竟」が共に出現頻度 1 で同一アラインメントに共起していたとしても，専門用語でない後者を候補語から除いておけば，前者は決定不能にならない。そのような見地から，AI 抄録において筆者らが専門用語と判断した名詞 1 単位に対象を限定して，これまで同様の調査を試みた。結果は表 11, 12 のようになった。表 11 から，専門用語に候補語を限定しても，決定不能状況に陥いる語は多いことが分かる。また，表 12 から，専門用語に限定した場合，先ほどの一般語も含めた場合よりさらに，カタカナ語・ローマ字語の除去が有効になることが分かる。

		AI 抄録		IP 抄録		IP タイトル		IP 文	
日本語	全異なり語数	4,422		16,354		10,417		16,354	
	頻度 1 の語	2,269	(51.31)	8,058	(49.27)	5,080	(48.77)	6,697	(40.95)
	頻度 1 で決定不能	2,031	(45.93)	5,658	(34.60)	1,505	(14.45)	2,645	(16.17)
英語	全異なり語数	4,151		16,629		14,050		16,629	
	頻度 1 の語	2,206	(53.14)	9,339	(56.16)	8,191	(58.30)	7,527	(45.26)
	頻度 1 で決定不能	1,961	(47.24)	6,813	(40.97)	2,767	(19.69)	2,651	(15.94)

表 4: 名詞 1 単位に関する決定不能状況

		AI 抄録		IP 抄録		IP タイトル	
日本語	全異なり語数	4,422		16,354		10,417	
	頻度 1 の語	2,269	(51.31)	8,058	(49.27)	5,080	(48.77)
	頻度 1 で決定不能	2,031	(45.93)	5,658	(34.60)	1,505	(14.45)
	(専門用語辞書除去後)	2,022	(45.73)	5,589	(34.18)	1,482	(14.23)
	(一般辞書除去後)	1,816	(41.07)	5,334	(32.62)	1,332	(12.79)
	(両方除去後)	1,811	(40.95)	5,295	(32.38)	1,327	(12.74)
英語	全異なり語数	4,151		16,629		14,050	
	頻度 1 の語	2,206	(53.14)	9,339	(56.16)	8,191	(58.30)
	頻度 1 で決定不能	1,961	(47.24)	6,813	(40.97)	2,767	(19.69)
	(専門用語辞書除去後)	1,950	(46.98)	6,768	(40.70)	2,731	(19.44)
	(一般辞書除去後)	1,790	(43.12)	6,569	(39.50)	2,639	(18.78)
	(両方除去後)	1,788	(43.07)	6,551	(39.40)	2,618	(18.63)

表 5: 名詞 1 単位に関する辞書対除外の効果

		AI 抄録	
日本語	延べ語数	43,395	
	異なり語数	2,240	
	頻度 1 の語	969	(43.26)
	頻度 1 で決定不能	703	(31.38)
英語	延べ語数	32,912	
	異なり語数	2,708	
	頻度 1 の語	1,226	(45.27)
	頻度 1 で決定不能	925	(34.16)

表 11: 名詞 1 単位の専門用語に関する決定不能状況

	AI 抄録	
異なり語数	2,240	
頻度 1 の語	969	(43.26)
頻度 1 で決定不能	703	(31.38)
(カタカナ語除去後)	406	(18.13)
(ローマ字語除去後)	295	(13.17)
(両方除去後)	64	(2.86)

表 12: 専門用語に関するカタカナ・ローマ字語除外効果

3 おわりに

対訳コーパスからの訳語対自動抽出では，低頻度訳語対の抽出が重要であるという前提のもと，本研究では頻度 1 語の調査を行い，以下の結果を得た。

(1) 決定不能状況に陥っている名詞 1 単位語は多い

(2) 辞書対の除外・カタカナ語の正規化・英語複数形の単数形化は，決定不能状況の改善にあまり有効でない

(3) カタカナ語・ローマ字語の除外は，決定不能状況改善に有効である

(4) 抽出対象を名詞 1 単位の専門用語に限定しても，上記と同様のことが言える

(5) 名詞の 2 単位連続の場合，決定不能語は非常に多く，カタカナ語・ローマ字語の除外処理も有効でなくなる

		AI 抄録		IP 抄録		IP タイトル		IP 文	
日本語	異なり語数	4,422		16,354		10,417		16,354	
	頻度 1 の語	2,269	(51.31)	8,058	(49.27)	5,080	(48.77)	6,697	(40.95)
	頻度 1 で決定不能	2,031	(45.93)	5,658	(34.60)	1,505	(14.45)	2,645	(16.17)
	(カタカナ正規化後)	1,909	(43.17)	4,976	(30.43)	1,337	(12.83)	2,207	(13.50)
英語	全異なり語数	4,151		16,629		14,050		16,629	
	頻度 1 の語	2,206	(53.14)	9,339	(56.16)	8,191	(58.30)	7,527	(45.26)
	頻度 1 で決定不能	1,961	(47.24)	6,813	(40.97)	2,767	(19.69)	2,651	(15.94)
	(単数形化後)	1,563	(37.65)	5,788	(34.81)	2,470	(17.58)	2,270	(13.65)

表 6: 名詞 1 単位に関するカタカナ語正規化・英語単数形化の効果

	AI 抄録		IP 抄録		IP タイトル		IP 文	
カタカナのみ	482	(21.24)	2,455	(30.47)	1,963	(38.64)	2,058	(30.73)
ローマ字のみ	710	(31.29)	3,305	(41.02)	1,705	(33.56)	2,663	(39.76)
ひらがなのみ	25	(1.10)	71	(0.88)	26	(0.51)	68	(1.02)
漢字のみ	944	(41.60)	1,936	(24.03)	1,257	(24.74)	1,649	(24.62)
その他	108	(4.76)	291	(3.61)	129	(2.54)	259	(3.87)
計	2,269	(100.00)	8,058	(100.00)	5,080	(100.00)	6,697	(100.00)

表 7: 頻度 1 の名詞 1 単位の構成字種

	AI 抄録		IP 抄録		IP タイトル		IP 文	
異なり語数	4,422		16,354		10,417		16,354	
頻度 1 の語	2,269	(51.31)	8,058	(49.27)	5,080	(48.77)	6,697	(40.95)
頻度 1 で決定不能	2,031	(45.93)	5,658	(34.60)	1,505	(14.45)	2,645	(16.17)
(カタカナ語除去後)	1,523	(34.44)	3,646	(22.29)	945	(9.07)	1,864	(11.40)
(ローマ字語除去後)	1,285	(29.06)	2,668	(16.31)	727	(6.98)	1,045	(6.39)
(両方除去後)	785	(17.75)	1,112	(6.80)	341	(3.27)	493	(3.01)

表 8: 名詞 1 単位に関するカタカナ語・ローマ字語除外の効果

		AI 抄録		IP 抄録		IP タイトル		IP 文	
日本語	延べ語数	13,047		112,276		68,091		112,276	
	異なり語数	5,802		36,554		25,975		36,554	
	頻度 1 の語	4,487	(77.34)	25,381	(69.43)	17,362	(66.84)	22,000	(60.18)
	頻度 1 で決定不能	4,372	(75.35)	23,656	(64.72)	9,642	(37.12)	12,568	(34.38)
英語	延べ語数	8,676		76,260		57,435		76,260	
	異なり語数	4,991		34,211		28,845		34,211	
	頻度 1 の語	4,154	(83.23)	25,897	(75.70)	21,635	(75.00)	23,311	(68.14)
	頻度 1 で決定不能	4,034	(80.83)	24,156	(70.61)	13,380	(46.39)	12,378	(36.18)

表 9: 名詞の 2 単位連続に関する決定不能状況

	AI 抄録		IP 抄録		IP タイトル		IP 文	
異なり語数	5,802		36,554		25,975		36,554	
頻度 1 の語	4,487	(77.34)	25,381	(69.43)	17,362	(66.84)	22,000	(60.18)
頻度 1 で決定不能	4,372	(75.35)	23,656	(64.72)	9,642	(37.12)	12,568	(34.38)
(カタカナ語除去後)	4,146	(71.46)	21,855	(59.79)	8,727	(33.60)	11,384	(31.14)
(ローマ字語除去後)	4,368	(75.28)	23,642	(64.68)	9,632	(37.08)	12,564	(34.37)
(両方除去後)	4,012	(69.15)	20,761	(56.80)	8,114	(31.24)	10,721	(29.33)

表 10: 名詞の 2 単位連続に関するカタカナ・ローマ字語除外効果

これらの結果から，次の点を指摘したい。即ち，対訳コーパスから低頻度訳語対を自動抽出する場合，少なくとも文をアラインメントとした統計的手法だけでは十分でない。アラインメントをより細かく設定する手法，あるいは翻字規則に基づくような他手法の併用を検討すべきである。特に複合語の場合は，固まりとして扱うのではなく，語構成要素同士の対応といった，より小さい単位での対訳性を考える必要がある。

今後は，実際の訳語対抽出実験を行い，上記方向の有効性を検証したい。

参考文献

[1] Gale, W. A. and Church, K. W.(1991) "Identifying Word Correspondences in Parallel Texts," *Proceedings of the DARPA Speech and Natural Language Workshop*, p.152-157.

[2] van der Eijk, P.(1993) "Automating the Acquisition of Bilingual Terminology," *Proceedings of the Sixth Conference of the European Chapter of the ACL*, p.113-119.

[3] Kupiec, J.(1993) "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora," *Proceedings of the 31st Annual Meeting of the ACL*, p.17-22.

[4] 熊野明・平川秀樹 (1994) "言語情報と統計情報を用いた対訳文書からの機械翻訳辞書作成," 自然言語処理, 100-12, 1994, p.89-96.

[5] Melamed, I. D.(1996) "Automatic Construction of Clean Broad-Coverage Translation Lexicons," *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, p.125-134.

[6] Smadja, F., McKeown, K. R. and Hatzivassiloglou, V.(1996) "Translating Collocations for Bilingual Lexicons: A Statistical Approach," *Computational Linguistics*, vol.22, no. 1, p.1-38.

[7] Vogel, S., Ney, H. and Tillmann, C.(1996) "HMM-Based Word Alignment in Statistical Translation," *COLING'96*, p.836-841.

[8] 大森久美子・堤純也・中西正和 (1996) "統計情報を用いた対訳単語辞書の作成," 情報処理学会第 53 回全国大会講演論文集, p.55-56.

[9] 高尾哲康・富士秀・松井くにお (1996) "対訳テキストコーパスからの対訳語情報の自動抽出," 自然言語処理, 115-8, 1996, p.51-58.

[10] 北村美穂子・松本裕治 (1997) "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, p.727-736.

[11] 佐藤健吾・中西正和 (1997) "最大エントロピー法による対訳単語対の抽出," 自然言語処理, 122-4, p.21-27.

[12] Ahrenberg, L., Andersson, M. and Merkel, M.(1998) "A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts," *COLING'98*, p.29-35.

[13] 米沢恵司・松本裕治 (1998) "漸進的対応付けによる対訳テキストからの翻訳表現の抽出" 言語処理学会第 4 回年次大会発表論文集, p.576-579.

[14] 大久保千英・坂井修一・田中英彦 (1999) "対訳テキストデータからの対訳表現の自動抽出," 情報処理学会第 58 回全国大会講演論文集, p.(2-273)-(2-274).

[15] Ker, S.J. and Chang, J.S.(1997) "A Class-Based Approach to Word Alignment," *Computational Linguistics*, vol.23, no.2, p.313-343.

[16] 山本由紀雄・坂本仁 (1993) "対訳コーパスを用いた専門用語対訳辞書の作成," 自然言語処理, 94-12, p.85-92.

[17] 石本浩之・長尾真 (1994) "対訳文書を利用した専門用語対訳辞書の自動作成：訳語対応における両立不可能性を考慮した手法について," 自然言語処理, 102-11, p.81-88.

[18] 松尾義博・白井諭 (1996) "発音情報を用いた訳語対の自動抽出," 自然言語処理, 116-15, p.101-106.

[19] Collier, N., Kumano, A. and Hirakawa, H.(1997) "Acquisition of English-Japanese Proper Nouns from Noisy-Parallel Newswire Articles Using KATAKANA Matching," *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, p.309-314.

[20] Breen,J.*EDICT* http://www.csse.monash.edu.au/ jwb/edict.html

[21] Shapiro, S. C. and Eckroth D.[編], 大須賀節雄 [監訳](1991)『人工知能大辞典』, 丸善, [New York: J.Wiley & Sons, 1987].

[22] コンピュータ用語辞典編集委員会編 (1996) 『英和コンピュータ用語大辞典第 2 版』, 日外アソシエーツ.