第三言語を介した対訳辞書の作成

田中（石井）久美子† 梅村恭司†† 岩崎英哉†††

第三言語を介して対訳辞書を自動作成するアルゴリズムを論じ，現存の対訳辞書との比較により抽出結果を検討する．和英辞典と英仏辞典から英語（第三言語）を仲介として和仏辞典を作成する際には，英語において機械的に辞書を合体させるだけでは第三言語の語の多義性により不適当な訳語が生じる．したがって，正しい訳語のみを自動抽出する手法が必要となる．本研究では，訳語関係のグラフ，語の形態素が持つ表意を利用した手法を提案する．実際に既存の中辞典規模の電子辞書を用いて名詞，動詞，形容詞に対して実験を行い，既存の辞書にはない訳語が得られるという有効性を示した．

Construction of Bilingual Dictionary Intermediated by a Third Language

KUMIKO TANAKA-ISHII,† KYOJI UMEMURA†† and HIDEYA IWASAKI†††

When putting two dictionaries, such as Japanese-English and English-French, together into a bilingual dictionary, the Japanese-French dictionary, it is necessary to discriminate equivalencies from inappropriate words caused by the ambiguity in the third language (which is English in the example). We propose a method to treat this by exploiting the structures of dictionaries to measure the similarity of the meanings of words. The resulting dictionary is a word-to-word bilingual dictionary of nouns, verbs, adjectives, and can be used to refine the entries and equivalencies in published bilingual dictionaries.

1. はじめに

調べたい見出し語や訳語が対訳辞書に記載されていない場合は，より国際性の高い第三言語を介して辞書を2冊引くことにより，望んでいた情報を得ることができることがある．第三言語を介して辞書を引くことは，国際性の低い言語や専門分野を扱う場合には不可欠である．この手法を自動化することによって，より多くの言語間，より多様な分野での対訳辞書を作成することが可能である．本論文の目的はその第一歩を目指すものである．

訳語を選別する手法としては，

- 基言語と第三言語間の辞書，第三言語と目的言語間の辞書を利用する方法
- コーパスなどの辞書以外の大規模データを利用する方法[1)]

が考えられる．両者を併用すれば，より品質の高い訳語関係を得ることができる．ところが，国際性の低い言語・専門性の高い分野などの場合には，コーパスそのものが存在しないなどの制約により，両者の併用が不可能であるような場面も多い．このような場合には，前者の方法のみに基づき訳語関係を獲得しなければならない．そこで，本論文では，前者の立場に立ち，辞書だけを用いて訳語を選別する方法を提案し，その可能性と限界を明らかにすることを目的とする．

本論文では，英語を介して日本語–仏語の対訳辞書を作成した．和英仏を選んだのは，

- 和英・英和，英仏・仏英辞典が電子的な形態で存在する，
- 仏語は英語に次ぐ国際語であるため，作成結果の和仏・仏和辞典を評価するための語彙が豊富で質の比較的高い仏和辞典が存在する，

という2つの理由による．

本論文の構成は次のとおり．2章では訳語を選択するためのアルゴリズムを説明する．3章ではアル

† 電子技術総合研究所
Electrotechnical Laboratory
†† 豊橋技術科学大学情報工学系
Department of Information and Computer Science, Toyohashi University of Technology
††† 東京大学大学院工学系
Faculty of Engineering, University of Tokyo

ゴリズムの性質をまとめる．4章では実験を行い，本手法の有効性を論じる．

和英，英和，英仏，仏英，和仏，仏和辞典を以下ではそれぞれ $\mathtt{Dic}_{\mathtt{J}\to\mathtt{E}}$，$\mathtt{Dic}_{\mathtt{E}\to\mathtt{J}}$，$\mathtt{Dic}_{\mathtt{E}\to\mathtt{F}}$，$\mathtt{Dic}_{\mathtt{F}\to\mathtt{E}}$，$\mathtt{Dic}_{\mathtt{J}\to\mathtt{F}}$，$\mathtt{Dic}_{\mathtt{F}\to\mathtt{J}}$ と記述する．また，$\mathtt{Dic}_{X\to Y}$ を $\mathtt{Dic}_{Y\to X}$ の逆辞書と呼ぶ．

注目する語に関して，日本語は**日本語**のような字体，英語は English のような字体，仏語は **français** のような字体を用いる．また，本論文では次のような表記法を用いる．一般に大文字は語の集合を，小文字は語を，イタリック体文字（*italic*）は変数を，タイプライタ文字（`typewriter`）は定数を表す．

- E，F，J はそれぞれ英語，仏語，日本語の単語の集合を表す．
- e，f，j は英語，仏語，日本語の単語を表す．
- X，Y，Z は，英語，仏語，日本語のいずれかを表す変数とする．
- x，y，z は，言語 X，Y，Z の単語を表す．

2. 問題とその解決方法

2.1 調和辞書

語を節，語と語の訳語対応を枝と見なすと，対訳辞書はグラフを構成する．枝には方向があるために，グラフは非対称な構造となる．$\mathtt{Dic}^{-1}_{X\to Y}$ をもって，$\mathtt{Dic}_{X\to Y}$ の枝の方向をすべて逆にした辞書を表すものと仮定する．

対訳辞書は同じ内容を指す語と語の間の対応関係を記述する[6]ので，語間の対応という観点からは枝が一方向でなければならない理由はない．したがって，一方向の枝をすべて双方向とし，このような辞書を言語 X と Y 間の調和辞書と呼ぶ．以後，調和辞書を D で，調和とは限らない一般の辞書は Dic で表記する．調和辞書においては $\mathtt{D}_{Y\to X}=\mathtt{D}^{-1}_{X\to Y}$ が成り立つ．以下提案する手法を適用する辞書は，すべて調和辞書とする．

2.2 逆引き法

ある日本語に対応する仏訳語を英語を介して得るための最も簡単な手法は，日本語の英訳語それぞれを機械的に $\mathtt{Dic}_{\mathtt{E}\to\mathtt{F}}$ で引くことである．たとえば，**図1** に示すように**競争**に対して得られる仏語は **compétition**，**concours**，**concurrence**，**race**，**course** などとなる．しかし，これらのうち **race** の意味は「人種」，**course** の意味は「競走」であるので，**競争**の訳語として適当ではない．機械的な対応により得られた仏語を訳語候補と呼ぶことにすると，訳語候補すべてをもとの日本語の訳語とすることはできないのである．

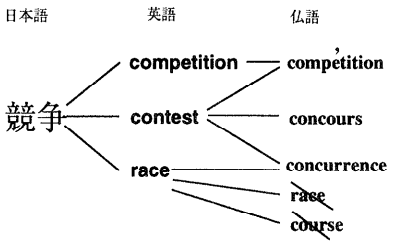

図1 「**競争**」の仏訳語候補
Fig. 1 Equivalence candidates for "**競争**".

race が訳語候補にあがった原因は，英語 race が同じ綴りで異なる意味（第一義は「競争」で，第二義は「人種」）を持つことである．訳語候補の **race** は「人種」の方の意味を直訳したために出現した．**course** に関しては，英語 race が日本語の**競争**より広い意味（すなわち，「急ぐこと」）を持ち，それを直訳したために生じた．

このような不適当な訳語が出現する理由として，以下の3点をあげることができる．

(1) 同型異義語（綴りが等しく意味が異なる語）を持つ英語を介した場合（**race** の場合）
(2) 日本語英語間，英語仏語間の対応する語に意味のずれがある場合（**course** の場合）
(3) 基とする辞書の訳語関係に間違いがある場合

訳語候補から適当な訳語を選別する従来の手法として，中間表現を用いた意味処理を考えることができる．中間表現を用いる場合は，日本語，仏語のすべての語に対して共通の人工言語で意味を記述し，訳語関係が正しいかどうかの判断はこの共通の人工言語上で行うこととなる．しかし，語の意味をどのように表現するのか，共通の人工言語として何を用いるのか，またその上でどのようにして語と語の意味の距離を定義するのか，共通の人工言語と日本語，仏語の対応をどのように構築するのか，などは大きな問題である．

より簡潔な手法としては，訳語関係を用いて訳語の適切度を表層的に測る手法が考えられる．その手がかりとして，我々が語学を学習する場合を考える．日本語に対する外国語訳語を探すときには，まず日本語 → 外国語の辞書を引き，外国語の訳語候補を得る．ところが，それらの語は外国語であるため，既知でないことが多い．したがって，外国語 → 日本語の辞書を各訳語ごとに引き直し（逆引きし），日本語の上で訳語の良し悪しを判断するのである．この操作は，異言語間の語の対応問題を同じ言語上の対応

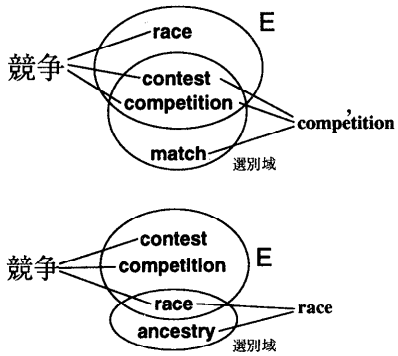

図2 1回逆引き法
Fig. 2 One time inverse consultation.

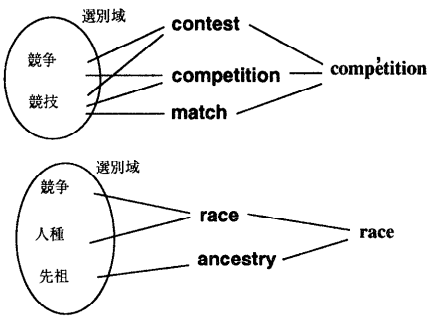

図3 2回逆引き法
Fig. 3 Two times inverse consultation.

問題に置き換えて判断可能にするという効果がある．

たとえば，**compétition**，**concours**，**race** をそれぞれ $\mathrm{Dic}_{\mathrm{F}\to\mathrm{J}}$ で逆引きすると，**競争**，**競争**，**人種** をそれぞれ第一義として得る．**人種** はもととなった日本語である**競争**とは何の関係もないので，**race** は訳語として不適当と判断することができる．このように，語を別の言語の語に対応させ，その語をもとの言語に再び対応させることによって不適当な訳語を削除する方法を，以下では「逆引き法」と呼ぶ．また，**race** に対する**人種**のように，訳語選別のために逆引きをして得られる語の集まりを「選別域」と呼ぶ．

本論文の三言語を扱う場合の最も簡単な逆引き法は，$\mathrm{Dic}_{\mathrm{F}\to\mathrm{E}}$ を用い，選別域を英語とするものである．図1の例では**図2** に示すように，各訳語候補を $\mathrm{Dic}_{\mathrm{F}\to\mathrm{E}}$ で引いた結果を選別域とし，それをもとの日本語に対応する英訳の集まり E={competition, contest, race} と比較する．たとえば **compétition** の選別域は {competition, contest, match} であり，E とは competition と contest が共通する．したがって，**compétition** を**競争**の訳語として選択する．一方 **race** の選別域は {race, ancestry} であり，E とは race しか共通するものがない．したがって，**race** は不適当な訳語と判断する．

本論文で扱う辞書はすべて調和辞書となるように前処理をする．そのため，選別域と E には少なくとも1語が必ず共通して含まれ，これは対象としている訳語候補が得られた際に経由された英単語である（図2下では race に相当する）．したがって，選別域と E に2語以上の語が共通するときに，訳語候補は正しいものと判断し，共通語が多い訳語候補ほど正しい訳語であるという立場をとる．この方法を1回逆引き法と呼ぶ．

1回逆引き法では選別域は英語であったが，さらに $\mathrm{Dic}_{\mathrm{E}\to\mathrm{J}}$ を $\mathrm{Dic}_{\mathrm{F}\to\mathrm{E}}$ の後に続けて用いると，選別域は日本語となる（**図3**）．この場合，**compétition** の選別域には**競争**が2つ，**競技**が3つ含まれる．表意文字「競」が含まれる数を考えると，訳語候補は**競争**と関係が深いと判断されるので，**compétition** は適当な訳語とする．一方，**race** の選別域には**競争**，**人種**，**先祖** がそれぞれ1回含まれる．**race** の選別域にはもとの見出し語**競争**に関連する語が**競争**以外に現れないため，訳語としては不適当と判断する．これを2回逆引き法と呼ぶ．

2.3 選別計算

各訳語候補に関する選別域を定めた後，2つの語集合を比較することによって，その訳語候補を選択するか否かを判断する．この処理を選別計算と呼ぶ．選別計算には表層的な字句に基づく方法と意味に基づく方法が考えられる．意味に基づく方法は，2.2節で述べたように，人工的な意味表現をどのように構築するか，また，意味表現の類似度をどのように測るかなどが問題となる．したがって本論文では前者を用いる．

選別計算には以下の4通りの方法が考えられる．第1に，2つの集合の共通集合に含まれる語数によって訳語候補を採択するという方法が考えられる．これは前節で1回逆引きを説明する際に用いたものである．

第2に，語に含まれる文字や文字列（字句）を利用する方法があげられる．選別域が日本語の場合には，字句として漢字を考えることができる．たとえば，**競争**の「競」や「争」が選別域の語に含まれるかを調べる．選別域に**競技**が含まれる場合は，**競争**と「競」の字が共通することを考慮し，訳語候補として採択する．これは前節の2回逆引きで用いている手法である．

漢字は表意文字なので，上の処理は表層的な字句を扱っているとはいえ，意味をも扱っていることにも相当する．英語や仏語では，接頭辞，接尾辞（たとえば英語で，international に含まれる inter など）の表意文字列を考慮することになる．形態素（語中において意味を有する最小言語要素）は意味を有するので，これを用いることにより，ある程度語の意味を処理することが可能である．

第3に類語辞典を利用することが考えられる．訳語を求める語を類語辞典で引き，得られた同義語が選別域に含まれる数を計測する．

第4に大量の電子テキストを用い，共起などの情報を利用することがあげられる．

本論文では，1章で述べたように対訳辞書データだけでどこまで処理が可能であるかを追求する観点から，第1と第2の手法を用いる．なお，第4の方法に関しては文献1)において論じている．

3. 逆引き法の性質

3.1 選別計算の定義

重み付き集合（multiset）とは集合の各要素が重みを持つ集合である．集合の要素を語，重みをその語が集合の中に重複して現れる回数によって定義する．たとえば，図3の **compétition** の選別域は重み付き集合で，要素**競争**の重みは2，**競技**の重みは3である．なお，1回逆引きによって得ることのできる訳語は単なる集合であるが，これはすべての要素の重みが1である重み付き集合と考える．

重み付き集合 X における要素 x の重みを式 $\delta_a(X, x)$ で表す．たとえば，図3の **compétition** の選別域を SA と記述すると，$\delta_a(SA, \textbf{競争}) = 2$ である．X が重み付き集合，Y が通常の集合のとき，$\delta_a(X, Y)$ を以下のように定義する．

$$\delta_a(X, Y) = \sum_{y \in Y} \delta_a(X, y)$$

Z={**競争**, **競技**}であるときは，$\delta_a(SA, Z) = 2 + 3 = 5$ となる．

さらに，式 $\delta_b(X, x)$ によって，重み付き集合 X の要素のうち，x と形態素を共有する語の重みの和を表現する．たとえば，$\delta_b(SA, \textbf{競争})$ は，5（「競」の字を含む選別域の要素の重みの和：**競争**の2＋**競技**の3）と2（「争」の字を含む選別域の要素の重みの和：**競争**の2）の和で7が得られる．さらに，Y を通常の集合とするとき，$\delta_b(X, Y)$ を以下のように定義する．

$$\delta_b(X, Y) = \sum_{y \in Y} \delta_b(X, y)$$

たとえば，$\delta_b(SA, Z) = 15$ であり，これは7に $8 (= \delta_b(SA, \textbf{競技}))$ を加えた結果である．δ_a, δ_b は，それぞれ前節で述べた選別計算の第1，第2の方法に対応している．

3.2 逆引き法の性質

ある日本語の語 j に対する仏訳語候補 F は重み付き集合であり，次のように記述される．

$$F = \mathrm{D}_{\mathrm{E}\to\mathrm{F}}(\mathrm{D}_{\mathrm{J}\to\mathrm{E}} j)$$

1回逆引きにおいて選別計算 δ_a を考えるとき，必ず次の条件を満たしている．

$$\forall f \in F \quad \delta_a(\mathrm{D}_{\mathrm{F}\to\mathrm{E}} f, \mathrm{D}_{\mathrm{J}\to\mathrm{E}} j) \geq 1$$

この理由は，調和辞書を用いているので訳語候補 f を得たときに経由した英語が必ず選別域に含まれるからである．同様に2回逆引きにおいても次の条件を満たしている．

$$\forall f \in F \quad \delta_a(\mathrm{D}_{\mathrm{E}\to\mathrm{J}}\mathrm{D}_{\mathrm{F}\to\mathrm{E}} f, j) \geq 1$$

この理由は，調和辞書を用いているのでもとの見出し語 j が必ず選別域には含まれるからである．

1回逆引きと2回逆引きに関して，次の性質が成り立つ．

性質1 $f \in \mathrm{D}_{\mathrm{E}\to\mathrm{F}}(\mathrm{D}_{\mathrm{J}\to\mathrm{E}} j)$ のとき，

$$\begin{aligned}\delta_a(\mathrm{D}_{\mathrm{E}\to\mathrm{J}}(\mathrm{D}_{\mathrm{F}\to\mathrm{E}} f), j) &= \delta_a(\mathrm{D}_{\mathrm{J}\to\mathrm{E}} j, \mathrm{D}_{\mathrm{F}\to\mathrm{E}} f) \\ &= \delta_a(\mathrm{D}_{\mathrm{E}\to\mathrm{F}}(\mathrm{D}_{\mathrm{J}\to\mathrm{E}} j), f)\end{aligned}$$

2つの補題を示したうえで，性質1を証明する．

補題1 $x \in \mathrm{D}_{Y\to X} y \Longleftrightarrow y \in \mathrm{D}_{X\to Y} x$

これは調和辞書の対称性から明らか．

補題2

X が集合（すべての要素が重み1）であるならば，

$$\delta_a(\mathrm{D}_{X\to Y} X, y) = \delta_a(X, \mathrm{D}_{Y\to X} y)$$

証明

$$\begin{aligned}&\delta_a(\mathrm{D}_{X\to Y} X, y) \\ &= |\{x | x \in X \wedge y \in \mathrm{D}_{X\to Y} x\}| \\ &= |\{x | x \in X \wedge x \in \mathrm{D}_{Y\to X} y\}| \quad \text{（補題1より）} \\ &= \delta_a(X, \mathrm{D}_{Y\to X} y)\end{aligned}$$

ただし，$|X|$ は集合 X の要素数を表すものとする．また，$\mathrm{D}_{Y\to X}\, y$ は枝が重複しないので，集合である．（証明終わり）

性質1の証明　枝が重複しないので，$\mathrm{D}_{\mathrm{F}\to\mathrm{E}} f$ は集合である．補題2より，

$$\begin{aligned}&\delta_a(\mathrm{D}_{\mathrm{E}\to\mathrm{J}}(\mathrm{D}_{\mathrm{F}\to\mathrm{E}} f), j) \\ &= \delta_a(\mathrm{D}_{\mathrm{F}\to\mathrm{E}} f, \mathrm{D}_{\mathrm{J}\to\mathrm{E}} j) \\ &= \delta_a(\mathrm{D}_{\mathrm{J}\to\mathrm{E}} j, \mathrm{D}_{\mathrm{F}\to\mathrm{E}} f)\end{aligned}$$

もう一方の等号に関しても同様．（証明終わり）

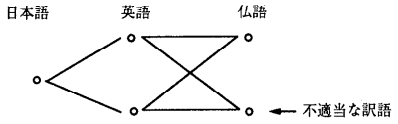

図4 1回逆引きが適用不可能な構造
Fig. 4 A structure that one time inverse consultation is inapplicable.

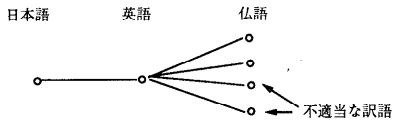

図5 逆引き法が適用不可能な構造
Fig. 5 A structure that our method is inapplicable.

この性質は，選別計算として δ_a を用いると1回逆引きと2回逆引きではまったく同じ結果となることを示している．しかも δ_a を適用した値は F 中の要素の重みと等しいことが分かる．すなわち，δ_a を選別計算とするのであれば，何も逆引きなど行わなくとも，F における重複数のみを用いて訳語が選択可能であるということである．

結局，逆引きの効果とは，三言語間のグラフの稠密度を利用して異なる言語の語の類似度をある特定の言語上に射影して測定するときに，和，英，仏のどの言語上でも測定を可能とすることである．

性質2 選別計算として δ_a を用いると，結果の和仏対訳辞書は調和辞書である．δ_b を用いた結果は必ずしも調和辞書とは限らない．

この性質は，δ_a を用いる場合は，対称な構造の辞書では明らかである．δ_b を用いる場合，結果の辞書の対称性は，形態素の用い方に依存する．

性質3 図4に示すような構造があり，一方の訳語候補を適当と，もう一方を不適当と判断するには，δ_b を選別計算として用いる必要がある．

図4の構造において，δ_a を用いると，選別域は2つの訳語候補に対してまったく同じものとなり，不適当な訳語候補を落とすのは不可能である．この場合は δ_b を用いることにより，片方だけ不適当と判定できる可能性がある．しかし，図5に示すように，ある見出し語に対する訳語候補が，1語の英語だけを経由する場合は，δ_b をもってしても不適当な訳語候補を落とすことができない．

4. 実　　験

4.1 辞書データ

実験に用いた辞書は和英[2]，英和[3]，英仏[4]，仏英[5]である．対訳辞書作成の全工程を図6に示す．まず，語と語の対応をそれぞれの辞書から抜き出す．品詞は名詞，動詞，形容詞に限定した．この作業はプログラムを用いて機械的に行ったが，辞書の記法は一貫していないものが多い．したがって，100語をランダムに選択して主観的に評価した結果によると，抽出結果には誤った対応関係が約18%も含まれてしまう．ここで得られた語対応を $\mathrm{Dic}_{J\to E}$, $\mathrm{Dic}_{E\to J}$, $\mathrm{Dic}_{E\to F}$, $\mathrm{Dic}_{F\to E}$ とする．

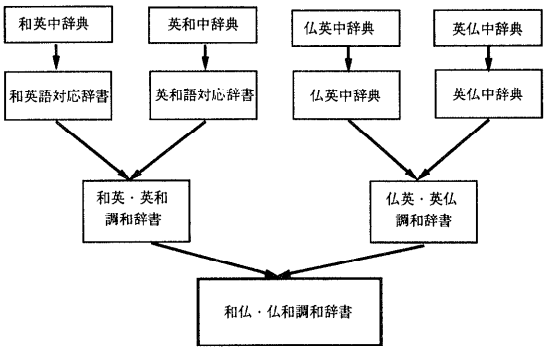

図6 対訳辞書作成の全工程
Fig. 6 Whole method.

次に，上で得られた語対応の辞書から調和辞書を以下のように作成する．

$$\mathrm{D}_{E\to J} = \mathrm{Dic}^{-1}_{J\to E} \cup \mathrm{Dic}_{E\to J}$$
$$\mathrm{D}_{J\to E} = \mathrm{D}^{-1}_{E\to J}$$
$$\mathrm{D}_{E\to F} = \mathrm{Dic}^{-1}_{F\to E} \cup \mathrm{Dic}_{E\to F}$$
$$\mathrm{D}_{F\to E} = \mathrm{D}^{-1}_{E\to F}$$

他にも調和辞書の作成方法は，たとえば双方向でない対応はすべて取り除く方法などが考えられるが，上のように処理したのは以下の辞書学上の理由[6]による．

対訳辞書は2種類存在する．第1は外国語から母国語への対訳辞書（$\mathrm{Dic}_{F\to M}$），第2は母国語から外国語への対訳辞書（$\mathrm{Dic}_{M\to F}$）である．ある外国語に対応する母国語が存在しない場合は，$\mathrm{Dic}_{F\to M}$ はその外国語の定義を母国語で説明文として与えることが可能であるので，$\mathrm{Dic}_{F\to M}$ の見出し語の範囲は外国語の語彙のすべてを覆う．一方，$\mathrm{Dic}_{M\to F}$ ではある母国語に対する外国語が存在しない場合は，その母国語そのものが辞書からの見出し語から削除されるため，$\mathrm{Dic}_{M\to F}$ で扱われる語彙範囲は母国語と外国語の共通する概念の語に限定される．本研究における調和辞書の作成方法は，このような偏りを解決する1つの方策である．

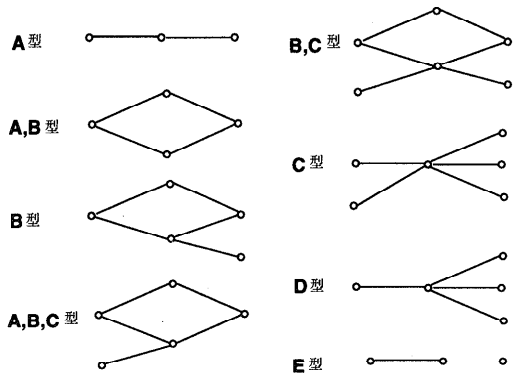

図7 見出し語の分類
Fig. 7 Classification of entries.

4.2 逆引きを用いた対訳辞書の作成工程

性質1より，δ_a を用いるときには1回逆引きと2回逆引きの結果は等しいので，δ_a を1回逆引きに用い，δ_b を2回逆引きに用いる．δ_b で用いる形態素は以下のように定義する．

- 日本語　漢字6353字
- 仏語　形態素1861文字列（接頭辞1151文字列と接尾辞710文字列）

性質2より，δ_b を用いる場合は結果が調和辞書とならない．したがって，逆引きは日本語を起点とするものと仏語を起点とするものと両方行い，2つの結果 $\mathrm{Dic}_{J\to F}$ と $\mathrm{Dic}_{F\to J}$ とを，

$$\mathrm{D}_{J\to F} = \mathrm{Dic}^{-1}_{F\to J} \cup \mathrm{Dic}_{J\to F}$$
$$\mathrm{D}_{F\to J} = \mathrm{D}^{-1}_{J\to F}$$

と，調和辞書に加工する．

起点の見出し語を，訳語候補の現れ方によって以下のように分類する（**図7**）．

- A型　ただ1つの訳語候補が存在する（介する英語は1以上）．
- B型　1回逆引きによって採択された訳語候補が存在する．訳語は，$\delta_a > 1$ を満たすときに採択する．
- C型　2回逆引きによって採択された訳語候補が存在する．訳語候補は，

  $\delta_b >$ (見出し語に含まれる形態素の数)

  を満たすときに採択する．
- D型　複数の訳語候補が存在するが，いずれも逆引き法によって採択されない．この型の見出し語は処理不能とする．
- E型　訳語候補がない．

正しい訳語が抽出される可能性は，A型の見出し語の方がB型のものよりも高く，また，B型の方がC型よりも高いと予想される．特に，B, C型は，選別計算時における δ_a, δ_b の値が高いほど，良い訳語と見なすことができる．

さらに，適用率を次のように定義する．

$$\text{全適用率} = \frac{\text{A, B, C型の見出し語数}}{\text{A, B, C, D型の見出し語数}}$$

$$\text{1回逆引き適用率} = \frac{\text{B型の見出し語数}}{\text{A, B, C, D型の見出し語数}}$$

$$\text{2回逆引き適用率} = \frac{\text{C型の見出し語数}}{\text{A, B, C, D型の見出し語数}}$$

C型の見出し語が同時にA型やB型であることもありうるし，B型の見出し語が同時にA型であることもありうるので，A, B, C型の見出し語は重複する．全適用率ではこのような重複を除いて率を計算する．

4.3 「競争」の処理結果

競争に対して採択された訳語候補は以下であった．

bague*, cage*, canal*, combat, compétition, concours, concurrence, course*, cuvette*, descendance*, émulation, joute, lignée*, lutte, race*, ras*, raz*, rivalité

介した英語は次の語である．

competition, contest, emulation, race, rivalry

なお，*印をつけた語は主観的に不適当と判断される訳語候補で，いずれもraceを経由して発生した不適当な候補である☆．

見出し語**競争**は1回逆引きによって選ばれた訳語が存在したのでB型であり，また，2回逆引きによって選択された訳語も存在したので，C型でもある．1回逆引きによって選択された訳語を次にあげる．ただし，括弧の中の値は，δ_a の値である．

compétition(3), **concours**(2), **concurrence**(3), **rivalité**(2), **émulation**(2)

raceから派生した不適当な訳語は落とされ，訳語候補の中からの選択としては妥当なものと判断される．既存の和仏辞書[9]では**émulation** を除いた4語の訳

☆ 電子辞書[4]に記載されていた英語名詞 race に対する訳語をあげておく．

race(1) f. 1. (in sea, river) raz m, ras m de courant. 2. Hyd.E: canal m, -aux. 3. Mec.E: ball r., (i) cuvette f, bague f, de roulement (pou billes); (ii) cage f à billes. 4. (a) Sp: course f; hundred metres r., course sur cent mètres; bicycle r., course de bicyclettes; horse r., course de chevaux; Turf: to go to the races, aller aux courses; r. meeting, réunion f de courses (de chevaux); (b) Pol: the arms r., la course aux armements.

race(3) f. 1. (a) the Mongolian r., la race mongole; r. riot, bagarre, émeute, raciale; (b) the human r., la race humaine. 2. (a) descendance f; of noble r., de sang noble; (of horse, dog, etc.) true to r., fortement racé; (b) lignée f.

語があげられていた．別の既存の和仏辞書[8]では，本結果に比べてさらに**compétition**が落ちている．これは本手法によって既存の辞書の品質を改善できる可能性を示している．

次に2回逆引きを用いたときの各訳語の δ_b の値をあげる．

bague(9)*, **cage**(8)*, **canal**(9)*, **combat**(16), **compétition**(21), **concours**(19), **concurrence**(21), **course**(25)*, **cuvette**(9)*, **descendance**(9)*, **émulation**(9), **joute**(9), **lignée**(8)*, **lutte**(18), **race**(8)*, **ras**(8)*, **raz**(8)*, **rivalité**(21)

競争の場合は，不適当な訳語候補はraceから派生する．和英辞典を逆引きする際にもraceにおける派生が起こり，日本語訳語として**競走**などがあがってしまう．したがって，すべての訳語候補に対して，選別域に「競」という共通の漢字が含まれ，δ_b の値は**競争**の文字数2よりも大きくなってしまう．一般に適当な訳語はすべて高い δ_b 値を示しているが，不適当な訳語は低い値となっている．したがって，δ_b に対しては何らかの閾値を導入する必要があることが分かる．

さらに，**émulation**や**joute**などといった適当な訳語は低い値を示している．また**course**は「競走」の意味であるので，不適当な訳語とされるにもかかわらず，最高の δ_b 値となってしまう．この原因は，**course**が日常的に使用される一般的な単語であるため，選別域が広がり，その中に多くの「競」の字が含まれているであるからである．**競争**から**競走**など，類似の意味へ派生する多義語raceを介するような場合は選別計算として形態素を用いる方法には限界があることが分かる．

4.4 見出し語の評価

表1に名詞，動詞，形容詞に関してA～Eの型別の見出し語の総数，適用率を日本語の見出し語，仏語の見出し語に分けて示した．この表から，以下のような傾向を読み取ることができる．

第1に，本結果は基とする辞書の品質に依存する．グラフの規模は英和間では大きいが，英仏間では小さい．これは表の総数の行の和仏の見出し語と仏和の見出し語の数の差として現れている．逆引きは起点が日本語仏語のどちらであっても同一のグラフを利用することには変わりがないので，逆引きの結果は規模の小さい英仏間のグラフに支配される．B型の行を$D_{J\to F}$と$D_{F\to J}$で比較すると，基とする2辞典の訳語関係の量に差がある場合でも，1回逆引きが適用された見出し語数は和仏，仏和でほぼ同じ規模となっていることが分かる．その分，1回逆引きが適用不能となった$D_{J\to F}$の見出し語は，A型，E型にまわり，A型やE型は$D_{F\to J}$に比べて多い．

表1 見出し語の評価
Table 1 Evaluation of entries.

	$D_{J\to F}$		
	名詞	動詞	形容詞
総数	46207	10129	15643
A型（候補数1）	9163	537	2313
B型（1回逆引き）	3981	2983	1061
C型（2回逆引き）	26888	6311	5450
D型（処理不能）	5233	2117	1299
E型（候補なし）	9516	477	7777
全適用率	75.8%	78.1%	83.5%
1回逆引き 適用率	10.9%	30.9%	13.5%
2回逆引き 適用率	73.3%	65.4%	69.3%
	$D_{F\to J}$		
	名詞	動詞	形容詞
総数	19433	4866	3788
A型（候補数1）	2631	318	214
B型（1回逆引き）	3449	2013	652
C型（2回逆引き）	3699	1522	656
D型（処理不能）	8546	1498	1307
E型（候補なし）	2626	402	1233
全適用率	49.2%	67.1%	48.4%
1回逆引き 適用率	20.5%	45.1%	25.5%
2回逆引き 適用率	22.0%	34.1%	25.7%

第2に適用率は2回逆引きの選別計算の能力に依存する．漢字と仏語の接頭辞と接尾辞の量には大きな差があるため，和仏に対してのみ適用率が高くなる．

第3に1回逆引きや2回逆引きの適用率は形容詞や動詞の方が一般に高い．多義性の高い品詞は訳語関係の分岐が多く，グラフが稠密になりやすい．すなわち，多義である品詞の方が英語での分岐が多くなり，1回逆引きや2回逆引きの適用率は高いと考えられる．その分，正解率が低くなることは次節で検証する．

D型，E型はA型からC型と傾向を異にする．D型に含まれるものは

- 英語1語を介し，多数の候補があがった見出し語
- 形態素が含まれない見出し語

の2通りに限定される．

E型の見出し語は，英語の見出し語が存在しなかったものである（図7参照）．以下にその主な分類と例を示す．

- 文化特有の語：**お年玉**，**cédille**

  文化特有の語は，英語において使用頻度の低い

語に訳されることが多く，特に同じような内容が目的言語の文化には存在しないときには，その見出し語は辞書には現れない．

- 専門用語，固有名詞：**ガウス，Cicéron**
  固有名詞や専門用語は，辞書が重点をおいた地域や分野に依存するため，英和，和英辞典と英仏，仏英辞典では記載される見出し語が異なる．たとえば，フランスの地名なども訳語候補が現れなかった．
- 借用語：**アペリティフ，tee-shirt**
  借用語は一貫しない綴りのために，英語において訳語と見出し語が一致せず，訳語候補が現れないものがあった（たとえば，apéritif は$D_{J\to E}$に英訳語として載っていた訳語である）．

調和辞書を用いた処理によって，現存の辞書にはない見出し語が処理結果に存在する．得られた$D_{J\to F}$を和仏辞典[8]と比較すると，4.1 節で述べた偏りが矯正される傾向が顕著に伺える．特に，

(1) 口語：**わんちゃん**
(2) 専門用語や固有名詞：**アスベスト**
(3) 日本語の複合語：**風化作用，物価安定政策**

といった見出し語が補填されていた．以上の結果は，母国語と外国語との対応を収集するために調和辞書に基づくことは有効であり，処理結果は現存の辞書の見直しや語彙の増加に役立つことを示している．

4.5 1回逆引きによって得られる訳語の評価

得られた訳語を各品詞のB型とC型の訳語関係（見出し語と訳語の組）の recall と precision を計算することにより評価した．

- recall　現存の辞書にその訳語関係が記載されている割合．
- precision　見出し語に対して主観的に正しいと判断される訳語の割合．

precision は主観的な判断であり，その割合が高いほど良い．recall は客観的な数値であり，現存の辞書と処理結果の辞書との訳語の一致度を表す．したがって，この値が低いことは現存の辞書を改訂する可能性を示している．

性質2より1回逆引きの結果は調和辞書である．したがって，和仏の訳語関係の組は仏和にも存在する．1回逆引きの評価は，仏和の訳語関係100組を任意に選んで既存の仏和辞書[7]と比較することにより，recall を計算する．結果を**表2**に示す．precision は，中間言語の多義性の度合に関係がある．英語は名詞よりも動詞の方が多義であり，多義であるほど precision が低くなると考えられる．実際，名詞であがった不適当な訳語は，類似の意味の訳語候補（たとえば**開花**に対し，**épanouissement**，**efflorescence** だけでなく，**bouquet**（花束），**fleur**（花）があがってしまったなど）にとどまっていたが，動詞であがった不適当な訳語にはまったく間違ったものが多く含まれていた．

表2 1回逆引きの結果に対する訳語の評価
Table 2 Evaluation of results obtained by one time inverse consultation.

	recall	precision
名詞	44%	76%
動詞	38%	60%
形容詞	15%	65%

表3 1回逆引きで得られた名詞の和仏訳語関係（δ_a の値の昇順上位50組）
Table 3 Noun translations of 50 highest δ_a values obtained by one time inverse consultation.

しみ	**tache**	停止	**arrêt**
傾斜	**inclinaison**	一致	**accord**
まぬけ	**idiot**	不足	**manque**
仲間	**camarade**	口論	**querelle**
刑務所	**prison**	ばか	**idiot**
利益	**avantage**	不一致	**désaccord**
評価	**estimation**	調和	**accord**
装置	**appareil**	絶頂	**apogée**
障害	**obstacle**	種類	**sorte**
口論	**dispute**	ばか	**imbécile**
おしゃべり	**bavardage**	変更	**changement**
変化	**changement**	平静	**calme**
物語	**récit**	評価	**évaluation**
評価	**appréciation**	注意	**attention**
増加	**augmentation**	制限	**restriction**
種類	**genre**	種類	**espèce**
子供	**gosse**	仕事	**travail**
光沢	**brillant**	刑務所	**taule**
傾斜	**pente**	強打	**coup**
下落	**baisse**	まぬけ	**niais**
まぬけ	**imbécile**	へり	**bord**
ばか者	**niais**	おしゃべり	**bavard**
騒音	**vacarme**	乱闘	**bagarre**
用心	**prévoyance**	勇気	**courage**
無資格	**incapacité**	妨害	**obstacle**

一回逆引きの選別計算で δ_a の値が大きい訳語ほど precision は上がる．実際，和仏辞典の名詞において δ_a の値の大きい上位50訳語関係に関しては precision はほぼ100%に近い．**表3**，**表4**，**表5** に各品詞別にこれを記載する．

recall は既存の辞書の改訂の可能性を示している．A☆，B，C型の結果の中で主観的に正しいと判断できる訳語関係から，その可能性のある訳語関係を分類すると，以下のようになる．

☆ A型の見出し語に対する抽出結果は多義性はもともと排除されている正しい訳語関係である．

表4 1回逆引きで得られた動詞の和仏訳語関係
(δ_a の値の昇順上位50組)
Table 4 Verb translations of 50 highest δ_a values obtained by one time inverse consultation.

投げる	lancer	静める	apaiser
やめる	abandonner	放棄する	renoncer
入れる	mettre	死ぬ	mourir
やわらげる	apaiser	放棄する	abandonner
悩ます	tourmenter	得る	prendre
投げる	jeter	置く	mettre
打つ	battre	静める	calmer
よごす	salir	やわらげる	calmer
なだめる	apaiser	無効にする	annuler
包む	envelopper	捕える	prendre
動かす	passer	盗む	voler
捨てる	abandonner	隠す	cacher
よごす	souiller	やわらげる	adoucir
だます	duper	ためらう	hésiter
理解する	comprendre	抑える	réprimer
抑える	étouffer	放つ	lancer
歩く	marcher	変える	changer
評価する	estimer	配置する	mettre
入れる	descendre	提出する	remettre
打ち倒す	abattre	静まる	calmer
飾る	orner	捨てる	renoncer
支持する	soutenir	刺激する	stimuler
広げる	étendre	穴をあける	percer
解く	défaire	引き起こす	produire
なだめる	calmer	だます	tromper

表5 1回逆引きで得られた形容詞の和仏訳語関係
(δ_a の値の昇順上位50組)
Table 5 Adjective translations of 50 highest δ_a values obtained by one time inverse consultation.

恐ろしい	affreux	正確な	juste
退屈な	ennuyeux	気まぐれな	capricieux
ひどい	affreux	ずるい	rusé
陽気な	enjoué	豊富な	abondant
ものすごい	affreux	明白な	manifeste
正しい	juste	重要な	important
厚かましい	effronté	強情な	entêté
ずるい	astucieux	勇敢な	vaillant
勇敢な	courageux	無限の	illimité
変わりやすい	inconstant	不十分な	insuffisant
不運な	malheureux	内気な	timide
適切な	juste	迅速な	rapide
気前のよい	généreux	気まぐれな	fantasque
陰気な	triste	まじめな	sérieux
すばらしい	épatant	すてきな	épatant
しっかりした	solide	けばけばしい	voyant
陽気な	joyeux	有害な	pernicieux
未熟な	inexpérimenté	法外な	exorbitant
変わりやすい	changeant	平凡な	banal
不当な	injuste	不思議な	merveilleux
不幸な	malheureux	熱烈な	passionné
内気な	modeste	注意深い	attentif
恥ずべき	honteux	怠惰な	paresseux
静かな	silencieux	生意気な	effronté
正当な	juste	上品な	élégant

- あまり使われない語
  cassine は既存の辞書には訳語見出し語そのものがなかった．本手法で得られた訳語は**掘っ建て小屋**であった．
- より自然な日常語への訳語
  adulation は既存の辞書には**過度の称賛，ほめすぎ，過褒，追従，へつらい**などとなっているが，本処理では**おべっか**なども訳語としてあがった．また，**bigarrè** は既存の辞書には**雑色**のという訳語しかないが，本手法により**ぶちの，まだらの**といった自然な日本語があがっている．
- 具体化された語の用法
  blanc は仏語では**白い**という形容詞であるが，その1つの名詞的用法として**卵白**が結果としてあがった．
- 和と仏の類似概念
  distique は仏文学の詩の用語であるが，本処理結果では日本語の同種の詩に対する用語**行連句，対句**が訳語としてあがった．一方，現存の辞書では直訳の**二行詩**を記載しているが，それは日本語としてはほとんどみない語である．ただし前者の方が分かりやすい反面，同等とはいえ別の文学用語を借用するという問題も内在する．

以上より，本論文で提案する処理は訳語の見直し，補填など辞書の改訂の補助として有効であると考えられる．

5. 関連研究

第三言語を介した対訳辞書の作成は，新たに辞書を手動で構築する際の常套手段である．たとえば，西和辞典[10]でも西英辞典の翻訳が基となっている．この辞書は，対スペイン語の中辞典の中ではその語彙の豊富さから現在日本では代表的な西和辞典である．手動では広く行われているにもかかわらず，自動化を試みた研究は他にはない．

辞書の自動構築ではないが，訳語関係のグラフ構造から概念項目を抽出する試み[11),12)]がなされている．これは，訳語関係のグラフが英和，和英辞典で訳語関係がサイクル，すなわち日本語の英訳語の日本語訳語がもとの語と一致する構造を成す場合に，そのサイクルが1つの概念項目を示すという仮定を検証するものであった．辞書のグラフのサイクルを，語の意味の近さを測る1つの尺度としている点で，本研究とは類似点がある．

本研究のもととなった論文13)では，1回逆引きと2回逆引き両方を同時に用いて訳語を選択し，評価している点が本論文とは異なる．2回逆引きでは δ_b の値に閾値を設けて，それを超えたものを訳語と定めており，それにより precision を上げることができている．

また，1章でふれた，辞書以外のデータを用いて訳語関係を得る研究として，対訳ではないコーパスを用いる方法を文献1)で論じている．この論文では，「第一言語で共起する2つの語の訳語は第二言語

でも共起する」という仮定に基づき，第一言語の共起情報を第二言語に翻訳することを確率行列の枠組みで形式化している．そのうえで，第一言語の共起情報の翻訳と，第二言語の基の共起情報が類似するように翻訳行列を最適化し，最適であるときは訳語関係の曖昧性が文脈に基づいて最も解消されている場合となることを示している．この研究の問題点は，最適な翻訳行列を求めるための収束計算時間が大きいため，ある程度訳語候補が絞りこまれたものに対してしか使用できない点である．したがって，辞書だけからある程度訳語を絞り込める手法と併用する必要性があり，本論文で提案する手法はその有力なものとなる．

6. 結　　論

第三言語を介し，辞書のグラフ構造を用いて対訳辞書を抽出する方法を提案した．本手法は第三言語の語の多義性から生じる不適当な訳語を排除し，正しい訳語のみを選択するために有用である．現存の辞書と比較して，結果の対訳辞書の品質を確かめ，既存の辞書の語彙の見直しや補填に役立つことが分かった．

本手法では，訳語関係のグラフを多義性の排除に用いるため，基辞書の品質，第三言語の多義性の度合が結果の精度に影響することなどが確認された．上の弱点を補うために，漢字などの形態素を用いた意味処理をも試みた．これは日本語など表意文字が多い言語には適用率の観点からは有利であるが，表音文字しかない仏語などの言語には不利となる．

参 考 文 献

1) Tanaka, K. and Iwasaki, H.: Extraction of Lexical Translations from Non-Aligned Copora, *Proc. 16th International Conference on Computational Linguistics* (1996).
2) 新和英辞典，研究社 (1990).
3) 新英和辞典，研究社 (1990).
4) *Shorter English-French Dictionary*, Harrap Limited (1982).
5) *Shorter French-English Dictionary*, Harrap Limited (1982).
6) Hartmann, R.R.K.: *Lexicography, Principles and Practice*, Academic Press (1983).
7) クラウン仏和辞典，三省堂 (1978).
8) スタンダード和仏辞典，大修館 (1970).
9) コンコルド仏和辞典，白水社 (1990).
10) 西和辞典，小学館 (1990).
11) 徳永，田中：対訳辞書からの概念項目の自動抽出，人工知能学会誌，Vol.6, No.2 (1991).
12) Tokunaga, T. and Tanaka, H.: The Automatic Extraction of Conceptual Items from Bilingual Dictionaries, *PRICAI* (1990).
13) Tanaka, K. and Umemura, K.: Construction of a Bilingual Dictionary Intermediated by a Third Language, *Proc. 15th International Conference on Computational Linguistics* (1994).

(平成 9 年 6 月 2 日受付)
(平成 10 年 3 月 6 日採録)

田中久美子（正会員）

1969 年生．1991 年東京大学工学部計数工学科卒業．1995 年仏国 CNRS-LIMSI 滞在．1997 年東京大学大学院工学系研究科情報工学専攻博士課程修了．工学博士．日本ソフトウェア科学会，ACM 各会員．

梅村　恭司（正会員）

1959 年生．1983 年東京大学大学院情報工学専攻修士課程修了，1983 年より 1995 年まで NTT 電気通信研究所勤務，1995 年より豊橋技術科学大学情報工学系助教授．東京大学博士（工学）．記号処理，プログラミング言語に関する研究に従事，ACM，ソフトウェア科学会，電子情報通信学会，計量国語学会各会員．

岩崎　英哉（正会員）

1960 年生．1983 年東京大学工学部計数工学科卒業．1988 年同大学大学院工学系研究科情報工学専攻博士課程修了．同年同大学計数工学科助手，1993 年同大学教育用計算機センター助教授，1996 年東京農工大学工学部電子情報工学科助教授を経て，1998 年東京大学大学院工学系研究科情報工学専攻助教授．工学博士．日本ソフトウェア科学会，ACM 各会員．