科学技術用語における2単語句の英日翻訳規則

4L－6

田上 文俊　竹田 正幸　松尾 文碩
九州大学大学院システム情報科学研究科

1. まえがき

著者らは，科学技術用語の英日対訳辞書に基づく単一語の訳語の抽出法を開発している[2, 3]．これを用いて辞書に未登録の用語の英日自動翻訳法を開発する計画である．翻訳は，単一語の訳語を組み合わせることによっておこなう．しかし，単一語の訳語は複数ある．この際に問題となるのが，複数ある訳語の組合せのうちどれを採用するかである．

例えば，contact point という見出しに対し，「接点」なる訳語が辞書[1]にある．contact の訳語には，辞書中の訳語として「コンタクト，混線，接触，接触面」，抽出訳として「巻付け，触発，接触型，接続，接」などがある．point の訳語は，辞書中の訳語「ねじ先，ポイント，航行，先，先端部，転てつ器，点，岬」，抽出訳として「端点，転てつ機，点エネルギー，点解析法」などがある．辞書の訳語「接点」は，contact の抽出訳「接」と，point の辞書中の訳語「点」を組み合わせれば得られる．本稿では，正しい組合せを得るための制約条件について考察した．

2. 単一語の訳語の抽出法

科学技術用語の対訳のデータとして，文献1(以後，科学技術用語辞書という) を用いた．この対訳辞書の英語見出しとその訳の対の集合を T で表し，英語見出しが単一語であるものの集合を T_0 で表す．T の異なり英単語数は 91,905，T_0 の異なり英単語数は 61,885 である．科学技術用語辞書 T には，

(1,2,3-trihydroxybenzene, ピロガロール)

のように，化学物質名が多数含まれていた．これらは特殊な訳し方をする．このため，抽出作業では，T から分野が化学，医学生物である対訳を除いた．これにより，T の異なり英単語数は 35,735，T_0 の異なり英単語数は 20,866 となった．

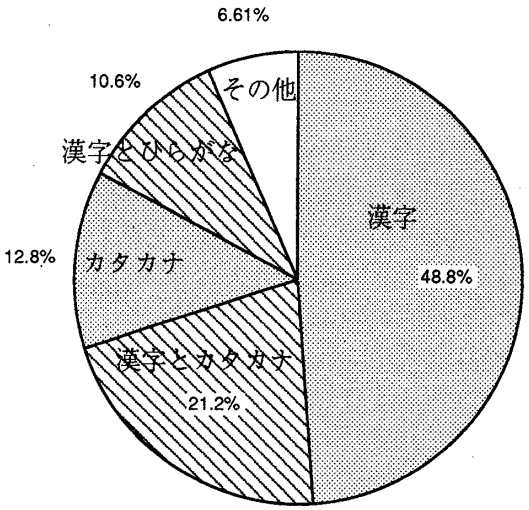

図1 2単語句の訳語の字種による統計 (総数 92,001)

単一語の訳語の抽出法については，詳しくは文献2, 3 を参照されたい．n 回目の抽出までに得られた対訳の集合を T_n とすると，T_5 で語数の増加は収束した．T_5 の異なり英単語数は 29,271 である．$T_5 - T_0$ の訳語を抽出訳という．T_0 のものを単に訳語という．

3. 字種による制約

字種による組合せの制約について考察する．まず，辞書中の英2単語句の訳語で用いられている字種を調査した．その結果を図1 に示す．訳語がカタカナと漢字からなる場合の漢字部分の文字数は，2 漢字である場合が一番多く全体の 57.7%，次に多いのが 1 漢字である場合で 18.1%であった．2 漢字については特に制約がなかったが，1 漢字については制約があるようである．例えば，contact の抽出訳「接」と point の訳語「ポイント」を組み合わせると，「接ポイント」という不自然な訳となる．

そこで，カタカナと組み合わせて用いられた 1 漢字

English-Japanese Translation Rules for Two-Word Phrase of Technical Terms
Fumitoshi Tanoue, Masayuki Takeda and Fumihiro Matsuo
Graduate School of Information Science and Electrical Engineering, Kyushu University, Hakozaki, Fukuoka, 812-81 Japan

がどのようなものであるかを調査した．2単語句の訳語がカタカナと1漢字からなり，そのうち1単語とカタカナ列が単一語の訳語で対応している場合について，残りの1漢字をそれぞれ調査した．

この1漢字が先頭語になる場合と最終語になる場合とでは傾向が異なることがわかった．先頭にしか現れなかった漢字は上位から'主','親','平','逆','丸'で，最終にしか現れなかった漢字は上位から'法','車','室','式','名'であった．先頭にきた漢字の異なりは225漢字，最終にきた漢字は200漢字，両方では354漢字が出現した．

4. 長さの上での制約

contact pointの「接点」という訳語は，単一語の訳語のうち最も短い同士を組み合わせたものと一致している．そこで，訳語候補をその長さによって選択することが考えられる．

一般に，短い訳語が好まれる傾向にある．長さによる制約を考える場合，カタカナ，ひらがな表記は長くなるので，ここでは漢字表記の訳語を考察の対象とした．漢字訳について，単一語の訳語を組み合わせたものがTにおける英2単語の訳語となる場合を抜きだし，そのうち，単一語の訳語がそれぞれ最短のものである場合と，さらにそのとき1漢字の訳語を使用した場合を抜きだした．それを集計した結果を表1に記す．

表1 単一語の訳語の連接が2単語句の訳語となる場合

	総数	最短な訳語を使用	1漢字訳を使用
辞書の訳語	8056	6056	2055
抽出訳使用	25709	9194	5201

辞書にあるT_0の訳語のみを用いた場合は，最短の訳語の組み合わせがそのまま2単語句の訳語となる場合が多いことがわかった．

カタカナ，ひらがな表記を含む場合については，それらの連続文字列を1漢字とみなした調査が必要であり，現在調査中である．

5. 1漢字の訳語の分類

長さの制約だけでは，最適な訳語の組合せを得ることはできない．ここでは，漢字同士を組み合わせる際の長さ以外の制約について考察する．

contact pointの例について，contactの抽出訳「接」とpointの訳語「端点」を組み合わせると，「接端点」という不自然な訳となる．そこで，2単語句の訳語が3漢字で，そのうち1単語と2漢字が単一語の訳語で対応している場合について，残りの1漢字がどのようなものかを調査した．

この1漢字が先頭語になる場合と最終語になる場合とでは傾向が異なることがわかった．先頭にしか現れなかった漢字は上位から'全','主','逆','横','半'で，最終にしか現れなかった漢字は上位から'機','率','器','計','系'であった．先頭にきた漢字の異なりは414漢字，最終にきた漢字は378漢字，両方では629漢字が出現した．

また，英単語の語順とその和訳部分の語順が異なる場合もある[3]．2字熟語について語順が入れ替わる場合には，次のようなものがあった．

(water intake, 取水) $\in T$

(water, 水) $\in T_0$

(intake, 取) $\in T_1$

上の例では，動詞の働きをもつ漢字が先頭に来ている．「取水」は，「水を取ること」を指し，「〜に(を)〜すること」という形で構成されている．この場合，動詞語義をもつ1漢字が，目的語の前に位置するため語順が入れ替わっている．そこでEDR日本語単語辞書を用い，動詞語義をもつ漢字を抽出した．その数は784である．

しかし，動詞語義をもつ漢字が2字熟語の後ろに位置する場合もある．同じ「取」という漢字を含む2字熟語には「油取」のような例もある．この2字熟語は「油を取るもの」を指し，「〜を〜するもの」という形で構成されている．この場合は語順は変化せず，前者との区別が必要となる．このため，動詞語義をもつ漢字を用いる場合，その漢字を抽出する際に語順が入れ替わったかどうかを，対応する英単語ごとに記しておく必要がある．

6. むすび

単一語の訳語を組み合わせる場合の制約条件を示した．1漢字の訳語を用いる場合は，多くの制約が生じる．それぞれの場合について，1漢字を分類した．

なお，本研究は，一部文部省科学研究費補助金(#07558162)の援助により行った．

参考文献

1) 日外アソシエーツ社：EB科学技術用語大辞典(1992).

2) 田上, 坂口, 竹田, 松尾：科学技術用語の英日翻訳規則の抽出, 情報処理学会第51回全国大会講演論文集 3-81 (1995).

3) 田上, 竹田, 松尾：科学技術用語の英日翻訳規則, 情報処理学会第52回全国大会講演論文集 3-89 (1996).