衛星放送用機械翻訳の辞書の改修
（慣用表現を中心に）

5Ｊ－4

畑田のぶ子　　浦谷則好　　江原暉将
（NHK 放送技術研究所）

１．はじめに

機械翻訳において、訳語選択は構文解析と並ぶ重要な要素技術である。筆者らはこれまで、辞書の改良という形で訳語選択の精度を向上させてきた。今回は「慣用表現」を利用した辞書の改良を行ったので報告する。慣用表現を利用することで、適切な日本語訳を出力することが可能である。慣用表現の収集には、BBI英和連語活用辞典（丸善発行、以下BBI共起辞書と呼ぶ）の使用権を得て利用した。BBI共起辞書から、見出し語と用例を含めて抽出し、抽出した名詞句、前置詞句、動詞句を分析して、新規に登録できるもののみを用いた。

改修の結果、基本辞書の語彙数は９９８５０語（新規登録２０８５０語）になり、ニュース専門辞書、翻訳者用辞書を合わせると約１５万語になった。

改修した辞書を用いて翻訳評価テストを行なった結果、全体の約５％の文が新規登録語を用いていた。その中で、翻訳結果が良くなったのが、６４％、ほとんど同じものが３３％、悪くなったものが４％であった。これより、全体では３％の文の翻訳の質が向上したことになり、慣用表現の登録が有効であることを確認した。本稿ではその概要を報告する。

２．収集データの分析

BBI共起辞書からの抽出データの内容を表１に示す。

表１：BBIから抽出したデータの内容

	不定冠詞 a, an		合計	割合
	含む	含まない		
名詞句	16033	7449	23482	55%
形容詞句	55	736	791	2%
副詞句	2	81	83	0%
動詞句	7864	6790	14654	34%
前置詞句	525	1764	2289	5%
その他	619	595	1214	3%
合計	25098	17415	42513	100%

基本的な品詞付与は下記の通りである。

名詞句：基本的に複合名詞

形容詞句と副詞句：それぞれ形容詞、副詞

前置詞句：副詞または叙述形容詞、後置形容詞

動詞句：現辞書で分析し、次のいづれかで登録する。

・全体を動詞（例：動詞＋目的語＝＞自動詞）として登録

・適当な部分に分割して登録。（動詞と前置詞句に分割、動詞と複合名詞、動詞と形容詞、etc）

採用するデータとしては、今回は、システム辞書に見出し語がないもの（新規登録語）に限定した。その結果、新規に追加された品詞は、大部分が名詞（９１％）となった。

名詞、形容詞、副詞、前置詞句はBBI共起辞書からのデータの７９％を利用できた。動詞句は、動詞だけでなく、名詞、形容詞、副詞、前置詞に分解され、利用できたものは少なく、約１０％にとどまった。

３．暫定版辞書作成と翻訳テスト

２.の新規登録用となったデータを、辞書に追加して暫定版辞書を作成し、翻訳テストを行なった。

・テストデータ：慣用表現は、通常の単語と比べると、出現頻度がかなり低いことが予想される。テストデータとしては５種のセット（合計６５９８文：表２参照）を用いた。

・評価方法：評価の対象とする文は第一訳に限り、訳語に新規登録語（新語）が使用されている文に限定

Improvement of NHK's MT Dictionary for News English - Idiomatic Expressions -

Nobuko Hatada, Noriyoshi Uratani, Terumasa Ehara　NHK Science & Technical Research Laboratories

した。文法の改修を行わなかったので、残りの文は訳の変化がなかったことになる。評価では、文全体が正しく翻訳されているかではなく、文のなかで使用されている慣用表現が、辞書の改修前と比べて良くなっているかどうかで評価した。

4．翻訳結果の分析と修正方法の検討

暫定版辞書で改修の影響を受けた文は全体の約５％で、そのうち、５４％は良くなり、１５％は悪くなっていた。問題になるのは、評価が低下しているものである。問題点とその対応について述べる。

1）訳語の問題

・訳語順、訳語の不足、特殊な訳語の問題については、新規追加の全訳語についてチェックすべきであるが、コストを考慮して、今回は、テスト文に現われた問題のあるもののみを調整した。

・翻訳結果から同じタイプの語に問題があると推測され、関連する語の分析が必要と思われるものに「単位の訳語」があった。そこで、出現する名詞や訳語をグループに分けて整理し登録した。

2）屈折形、および屈折形を含む句を見出し語にするときの問題、

・現システム辞書の見出し語の屈折形が見出し語として新規に登録されることによる問題

・現在分詞／過去分詞＋名詞を複合名詞で登録する場合の訳語の問題

があった。問題の有無を調べて優先順位を調整した。

3）句を一つの品詞で登録することにより、句に含まれるべき句の前後の修飾部分が句の外の別の語を修飾してしまう問題（例えば、名詞句の形容詞にかかるべき副詞が動詞にかかってしまう）。

問題の有無を調べて、今回登録した語を用いない解も上位にくるように優先順位を調整した。

5．修正データによる辞書作成と翻訳テスト

辞書修正後の評価結果を表２に示す。「良い」が１０％増加し、「悪い」は１０％減少し、調整がうまくいったことが分かった。以上より、調整後の辞書を用いた翻訳では、テスト文の５％が新規登録語を用いているので、全体では３％の文の翻訳の質が向上したことになる。

表２：評価結果（修正後）

テスト文	文の数	新語 使用 文の数	新語 使用 個数	個数に 対する 割合 ○	△	×
AP	180	20	20	55%	35%	10%
JD	310	7	7	71%	14%	14%
NRT_ALL	1763	45	47	66%	28%	7%
bs.1994.1-4	2245	148	157	70%	28%	2%
TEST_SEN	2100	86	87	56%	41%	2%
合計	6598	306	318	65%	32%	4%
暫定版	6598	318	328	55%	32%	14%

○：良い、△：同じ、×：悪い

6．まとめ

評価で「良くなった」ものは、その文に対してその訳語が適当であったことを示しており、文が異なるとその訳語が適当でない場合もあることに注意する必要がある。反対に、「悪くなった」と判定された文でも、訳語選択に依存するものは、文が異なれば、その第一訳語が適当となる場合もある。今回の結果は慣用表現を使用しても、翻訳の精度を上げるには、共起する動詞や名詞の意味、文脈をどのように整理し、扱うかという基本的な問題が依然として残ることを示している。

今回のように、慣用表現を１語で登録することで生じる問題が２つある。１つは、句を一つの品詞で登録することにより、句に含まれるべき句の前後の修飾部分が句の外の他の語を修飾してしまう問題である。もう１つは１語で登録することによる柔軟性のなさである。これらは、ローカルな処理の拡張で解決可能と考えるが、慣用表現をどのように整理分類し、処理するかが、これからの課題である。

参考文献　畑田、田中、江原、浦谷、加藤：ニュース用英日機械翻訳システムの課題と改善～辞書、文法とその改修を中心に～、NHK技研R&D、No４５、１９９７年、５月、ｐ１～２２