自然言語処理 121－14
情報学基礎 47－6
（1997. 9. 12）

対訳付慣用表現の収集について

田中康仁
兵庫大学
E－mail : yasuhito@humans－kc. hyogo－dai. ac. jp

概要

対訳付慣用表現の基礎となるデータの収集方法、ファイルの作成、慣用表現の標準化の問題点、訳語の問題点について述べている。さらに慣用表現例文ファイルのデータ収集、検索方法について述べ、研究の基礎的データの作成と検索方法を分析した。

Collecting Idioms and Their Equivalents

Yasuhito Tanaka
Hyogo University
2301 Shinzaike Hiraoka Kakogawa
Hyougo Japan 675－01
Email:yasuhito@humans-kc.hyogo-dai.ac.jp
TEL +81－794－27－5111 FAX +81－794－27－5112

Abstract

This paper details the method I used to collect the data that formed the basis of idioms and their equivalents, the data filing system, and the problems I experienced in standardizing and translating the expressions. It also examines how I gathered the data for illustrating the general use of the idioms and the method I used to seach those data, and analyzes the way I drew up and searched the basic data used in the research.

0）はじめに

機械翻訳システムが多くのメーカーやソフトウエア会社から発表されている。しかし、品質はまだまだまだであり、充分使えるというほどのものではない。この理由としては、「文は単語の集まりである。」という考えから単語辞書は充実させたが、それ以外の単語と単語の結合したものについてはあまり充実していない。専門用語については少し集められている。

一方、「文は類似した文の学習から作られる。」という考えに基づき、文の類似度からの翻訳が研究されて実用化なされようとしている。これにも、どれだけの文の集合でよいのか?, どれだけの類似度で類似とみなすかなど問題は多い。そこで、筆者は単語と単語の結合したものにもっと関心をはらうべきだと考えている。

機械翻訳の基礎的知識（Basic Knowledge）としては単語、複合語、専門用語、概念、概念と概念の関係、慣用表現等が考えられる。個々に研究を進めなければならないが、私の考えでは複合語と慣用表現の対訳付データの大量の収集と研究、体系化がなかなか進んでいないと考える。これは機械翻訳の応用分野がある特殊な分野に限られているためで、広くて一般的な分野に移れば、この分野は重要な問題になると考えている。数量としても数万から数十万のデータが必要である。

また高級な知識とは：「水は摂氏℃0度で凍る。」とか「アメリカは原爆を日本の広島と長崎に落とした。」「その結果、数十万の人々を殺した。」というようなものである。高級な知識又は基礎的な知識の中間に位置するようなものも考えてゆかねばならない。例えば諺等がそれである。しかし、ここでは基礎的知識の対訳付データについて考える。

〔1〕慣用表現について

1.1）慣用表現の数量的分析

慣用表現（熟語）がどの程度使われているか数量的に分析した資料を調べてみた。旺文社刊の「英熟語ターゲット1000」には16年間の日本全国の大学入試に出された試験問題から慣用表現について調べている。その異なりの総数は14,368である。

1回	……	8,083	（56%）
2回	……	1,884	（13%）
3回	……	926	（6.4%）
4回	……	532	（3.7%）
5回	……	417	（2.9%）
		11,842	（82%）

出典「英塾語ターゲット1000」旺文社

大学入試はおもしろい対象である。これで全ての慣用表現を推定するのは大変であるが一つの目やすになる。

筆者の考えでは、この4〜5倍程度は集めなければならないと思う。5倍とすると7万程度の慣用表現を集めなければならないことになる。

1.2）慣用表現をどのように考えるか

慣用表現を集めた本を数冊調べその索引を見ると、次のようになっている。

（i）動詞句 ex. 動詞⊕前置詞のようなもの
　　例：look up
（ii）前置詞句
　　例：in front of
（iii）副詞句
　　例：as soon as
（iv）be動詞⊕動詞句
（v）否定形notを含むもの
（vi）その他

がある。このうち（i）、（ii）が多い。

1.3）英語の前置詞

英語の前置詞を調べてみると3種類のものがある。これは首藤訓宏、中島武著（「前置詞・イディオムの英会話」－重要20前置詞で表現自在－ダイヤモンド社）から引用する。

前置詞を分類すると次のようになる。

1．Simple Prepositions「単純前置詞」
2．Compound Prepositions「複合前置詞」
3．Participial Prepositions「分詞前置詞」

これらは次のようになっている。

Simple Prepositions（23）

at	forth	on	till
but	from	out	to
by	in	over	up
down	like	per	via
* ere	of	save	with
for	off	through	

（ere[ɛ]「前に」の意だが，現在では詩などに使われるだけ）

Compound Prepositions (43)

abaft	amidst	besides	throughout
aboard	among	between	toward (s)
about	* anent	* betwixt	under
above	around	beyond	underneath
across	* aslant	despite	until
* adown	* athwart	except	unto
after	before	inside	upon
against	behind	into	* withal
* aloft	below	onto	within
along	beneath	outside	without
amid	beside	since	

Participial Prepositions (13)

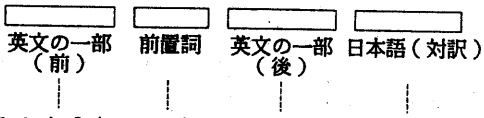
* barring	during	past	respecting
bating	excepting	pending	saving
concerning	* notwithstanding	regarding	touching
considering			

(* barring＝excepting「を除いて」。notwithstanding「にもかかわらず」)

これらのうち単純前置詞だけを調べればほぼ良いことが前述の旺文社の資料を調べる中でわかる。

1.4) 英語の前置詞に注目する。

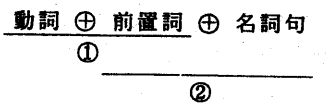
英文と日本語文が対になったパラレル・コーパスを分析の対象と考える。しかし、パラレル・コーパスはなかなか入手できない。もし、これらが入手できるとすると、英文を中心にして考える。まずKWICを作成する。次のようなものを作る。



このようなKWICを作成し、前置詞を含む右側で分類すると前置詞句が集めやすいKWICになる。

また、前置詞を含む左側にむかって分類すると動詞句が調べやすいことがわかる。人手による調査、エディターによる調査、自動（半自動）抽出等について考えねばならない。

英文の中には次のような構成になっている場合がある。



一つの可能性として動詞と前置詞が結合し、動詞句を作っているのか、前置詞と名詞句が結合して前置詞句が出来ているのか、はっきりとしない場合がある。このような場合には人間の援助が必要である。

1.5) コーパスからのKWICの作成について

Wall Street Journal のCD－ROMを使用して前置詞のKWICを作成した。この前置詞の前の語で分類したものと後の語で分類したものを作成した。これらKWICは動詞句の分析、前置詞の分析に有効であることがわかった。

これらの前置詞のKWICの件数は次の通りである。

(1) in	634,857	(7) but	24,325
(2) for	325,876	(8) out	21,088
(3) by	195,710	(9) over	18,299
(4) on	177,050	(10) per	9,542
(5) from	138,031	(11) off	4,355
(6) at	136,823	(12) down	3,877

このようになっている。特にin, for, by, on, from, atの6つが特に重要である。

慣用表現を抽出するには対訳コーパスを使い抽出する方法が考えられる。

もう一つの方法は辞書からの抽出である。

5) どのような辞書や本を対象とするか？

厚い本で、立派な慣用表現の辞書がある。しかし、これら本は良いものであるが次のような問題がある。

i) 著作権の問題

ii) 本ができるまでに長期間かかっているため少し、古くなっている。

iii) 本の著者によって慣用表現の選別と訳語の選別がなされている。これは良い面もあるが、悪い面もある。

本からの引用には著作権の問題が常につきまとうことは忘れることはできない。

次のような本や雑誌を対象とする。

高等学校の生徒が使用する参考書、雑誌、How to 物のような実用書の中から慣用表現（英語）と対訳を頭の中に入れ記憶し、その後、その内容を抽出する。

このような本は、厚い本を作成する前段階として作られたものや、商業主義的なおもしろさや実用主義を中心としたものである。しかし、

英語という巨大な対象物を完全に調べつくすことは簡単ではない。そこで、多くの人の少しずつの知識をうまく集め、体系化すればよいと考える。

例文は抽出しない。

例文は別の方法で作成したコーパスの中から抽出し、訳語の検証等に利用する。

〔２〕慣用表現の収集

２.１)どのように集め入力するか、入力形式について述べる。

1) No: 番号
2) 英文慣用句
3) 和訳（複数可能）
4) 書籍記号
5) 頁番号:該当する頁番号

これらは逆スラッシュ（＼）をもちいて区切りを行う。次のように入力した。

例
＼1＼05515
＼2＼be unable to ［A］
＼3＼［A］することができない
＼4＼F
＼5＼12

＼1＼05502
＼2＼call at
＼3＼訪ねる
＼4＼F
＼5＼10

２.２）入力データ件数

14冊の本から慣用表現を抽出し入力、整理した。１つの英語の慣用表現にどの程度の訳語が付しているか調べてみると次のような結果になった。

1 種類	9,244 件	8 種類	20 件
2 種類	6,087 件	9 種類	13 件
3 種類	1,824 件	10 種類	9 件
4 種類	523 件	11 種類	3 件
5 種類	173 件	12 種類	1 件
6 種類	74 件	13 種類	2 件
7 種類	37 件	合計	18,010 件

これらについてはさらに内容を整理、検討している。また、さらに８冊の本を追加し、約３万件のデータとなっている。

２.３）慣用表現の標準化

慣用表現は文の中から抽出し、次のような標準化を行わなければならない。

1) 動詞の過去形は原形にする。
was able to→be able to
is ,was , were……→be
2) 所有形について
in his wayhome→in one's wayhome
かえり道
3) 再帰代名詞→oneself
dried himself off →dry oneself off
体をかわす
4) 複数形は単数形にする
soft contact lenses→soft contact lense
5) Theはa にする。
しかし、特殊なものはそのままにする。
例 The Japanese a Japanese
国民全体 個人
6) 短縮形は短縮でないものにする。
don't → do not
I've → I have
7) 語をまたがるものについては［A］［B］等の記号を入れる。
launch［A］in［B］←→［A］を［B］に送り出す。
8) 動詞＋ing→動詞
driving wasps off→drive wasp off
はちを追い払う
9) someone,something（誰かを表す、何かを表す）等を使って抽象化されているものもある。
10) doing⇒動詞の進行形を使うべきところにdoingを用いて抽象化している場合がある。
11) その他

標準化はあまり行いすぎると一般文に適用する際にむつかしくなる。

２.４）慣用表現と訳しわけ

慣用表現には色々な訳語が付いている。

例 best season: 最も良い時、良い時期

しかし、実際には次のように使われている。

The cherry is best season.
桜は満開である。

このため、慣用表現の訳語としては代表的な訳語を採用している。

しかし、実際には代表的な訳（複数の場合有り）がどの程度使われているか、異訳(表記上のちよっとした場合と、異なった意味の場合がある。)がどのように使われているか、その条件は何か等を調べてゆかなければならない。

英語の動詞句は日本語の複合動詞と似た面がある。それ故動詞句には曖昧さが付きまとう。このため使用条件を明確にしなければならない。

［３］慣用表現を含む例文の収集について

慣用表現を含む例文をどの程度集めればよいか考えてみる。

・慣用表現を約5万件集めるとする。
・1つの慣用表現に平均5つの例文を対応させるとする。

5万件 × 5 ＝ 25万件（例文）

約25万件の対訳付例文を集める必要がある。これは大変な作業であるが実行しなければならない。

[4]慣用表現の例文と検索について

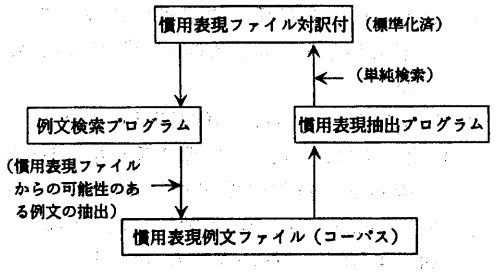
慣用表現とその例文は慣用表現の本等では対応しているのが普通であるが、ここでは別ファイルとして管理することを提案したい。
対訳付例文ファイルは慣用表現ばかりでなく次のような目的にも巾広く利用できる。

(1)機械翻訳用テストデータとして利用する。
(2)文型パターンの抽出例文とする。
(3)例文による翻訳システムの開発の基礎資料となる。
(4)慣用表現の訳し分けの材料となる。
(5)その他

等があげられる。

また対訳付データを年々増加させていくためには例文と慣用表現ファイルとが別のファイルとしておくことが望ましい。

慣用表現ファイルと慣用表現例文ファイルとの関係を示すと次のようになる。



慣用表現抽出プログラムについては既に前回述べているのでここではふれないことにする。

例文検索プログラムは慣用表現（標準化済）ファイルから現実に使われている表現に合った形になおし検索を行うものである。現実に使われている表現になおす例としては、動詞は各変化形になおす。be動詞は（is, was, are,were…）に変える。単数は複数形に変えることもあるといったものである。

それ故、例文検索プログラム、慣用表現抽出プログラムは次のテーブル又は辞書を持っていなければならない。

(1)代名詞の変化表
(2)不規則動詞の変化表
(3)辞書　名詞の単数、複数形
　　規則動詞の語幹、語尾の変化型

これらプログラムは慣用表現の抽出、検索と同時に例文の頻度集計も行わなければならない。

例文検索プログラムでは英語の表現が抽出されても、日本語の訳語(複数個のもの)が一致するものが各々どれだけあるか、全く異なった訳語がどれだけあるか調べなければならない。また異なった訳語が発生する要因は何かも分析しなければならない。

また、標準化されているため慣用表現例文ファイルにあるにもかかわらず検索することができないものがあるかもしれない。これらについては単語(原形)の使用頻度を調べることや検索する語を原形や標準形に変え、慣用表現例文ファイルの構成単語の原形や標準形の論理積による検索により、抜けを監視し、プログラムの改善に勤めなければならない。

[5]慣用表現用例文ファイル

慣用表現用例文ファイルは日本語と英語の対になったもので、その中に慣用表現が使われていればよいものである。慣用表現が使われていなくても特にファイルとして問題になることはない。

次の項目が必要である。

1) No. 番号
2) 日本語文
3) 英文
4) 参考文献（記号）
5) page No.
6) 検索キーワード

例 ＼1＼ 0001
　＼2＼ 彼女は君の野心に軽蔑しているぜ。
　＼3＼ She is contemptuous of your ambition.
　＼4＼ B 000
　＼5＼ 27
　＼1＼ 0010
　＼2＼ 彼女は犯罪に巻き込まれた。
　＼3＼ She was involved in the crime.
　＼4＼ B 000
　＼5＼ 97

このはか計算機処理、又は、一部手作業で英語の慣用表現の検索キーを項目として持つことも考えている。これは計算機処理の効率化のためである。

このように逆スラッシュで区分記号を付けていると処理しやすい。

〔6〕慣用表現例文ファイルの作成方法

1) 人手による入力

1人5,000例文を入力するとし、10人で1年間に約5万例文の入力が可能である。これは実現可能である。5年たつと約25万例文のコーパスが作成できる。

2) CD－ROMや電子ブック（ディジタルブック）の利用

CD－ROM化されている英和・和英辞書を編集しなおすことも1つの方法である。

また、最近は電子ブック(ディジタルブック)が安く入手できる。これをパーソナルコンピュータで読ませることが可能である。このようなソフトウエアも電子ブックとして市販されている。

電子ブックやディジタルブックの内容をテキストファイルとして利用できる市販ソフトウエアとしては、「電子ブックをパソコンで読もう!」電子ブック検索ソフト付 ㈱メディアパル 3,980円

フリーソフト等を利用することにより、個人の研究や翻訳に利用できるデータが数多く得られる。

電子ブックを単体で利用するハードウエアは、ソニー社からデータディスクマンとして売られている。

また、電子ブック、ディジタルブックは紀ノ国屋書店で多くが取扱われている。これらは本屋の流通経路で売られている。約300ほどのものがある。今後はさらに拡大するようである。

3) 本や雑誌をOCRで読ませ編集する。

本や雑誌の一部をOCRで読ませ編集することも考えられる。

一昔前はOCRの精度、文字認識の精度、誤り抽出、訂正の方法もあまり良いものでなかったが、今ではこれらの技術が向上し、総合的に実用に耐え得るものにまでなった。

但し、常に著作権に注意しなければならない。個人の研究利用やテスト・データとしては利用可能でも、商品開発においては問題である。

〔7〕慣用表現収集についての基礎的な検討ができた。今後はさらに応用するにあたっての問題点や訳語選択等のことも考えてゆかなければならない。

言語の研究を行うには慣用表現であるとか慣用表現用例文ファイルといった基礎的資料を充実させ、着実に進まなければならない。科学、工学として言語を取り扱うためには、データの準備が重要である。

対訳付慣用表現の問題点を考えながらデータの収集を開始した。今後このデータが機械翻訳システムの品質向上に大きな役割をはたすことを期待している。

「量的変化は質的変化をもたらす」という哲学の言葉を信じ、実行してゆきたい。

〔8〕参考文献

1) 花木金吾「英熟語ターゲット1000」旺文社 ISBN 4－01－030754－4

2) 首藤訓宏、中島武「前置詞・イディオムの英会話」－重要20前置詞で表現自在－ダイヤモンド社 ISBN 4－478－98011

3) 田中康仁、吉田将 慣用表現について －収集と整理－情報学基礎5-1 情報処理学会 1987. 6

4) 田中康仁 吉田将 概念辞書の作成 自然言語処理75-12 情報処理学会 1990.1

5) 田中康仁 対訳付慣用表現の収集について 第52回(平成8年前期)全国大会講演論文集(3) 情報処理学会 4B－9 1996. 3 pp3－67, 68

6) 田中康仁 対訳付慣用表現の収集について(その2) 第53回(平成8年後期)全国大会講演論文集(3) 情報処理学会 4L－03 1996.10

7)「電子ブックをパソコンで読もう!」電子ブック検索ソフト付 ㈱メディアパル